A feedback system for a children's helpline training-chatbot

A. A. Braam

TU Delft

A feedback system for a children's helpline training-chatbot

by

A. A. Braam

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday December 12, 2023 at 09:00 AM.

Student number: 4456327
Project duration: September 20, 2022 – December 12, 2023
Thesis committee:

Dr. ir. Willem-Paul. Brinkman,	TU Delft, supervisor
Ir. Mohammed Al Owwayed,	TU Delft, daily supervisor
Dr. ir. Ujwal Gadiraju,	TU Delft, committee member

An electronic version of this thesis is available at http://repository.tudelft.nl/.

TU Delft

Acknowledgements

Where to begin... it has been a long, arduous trek but we have made it. I am quite frankly at a loss for words so I will just take this one page and thank everyone who guided and supported me throughout this whole process. I do believe that the first thanks should go to Mohammed Al Owayyed, whose kindness and expertise made this experience fun and possible. I also want to thank Willem-Paul Brinkman for offering much-needed structure and guidance when it was most needed, and Nele Albers for her insights and counsel. A special thank you also goes out to Martin Dierikx and Andrei Stefan, whose friendship and work ethic both entertained me, and kept me in line when I was losing focus. I would also be amiss if I did not mention Dongxu Lu, whose contributions were invaluable to this project. Finally, I want to give a shout-out to everyone else who helped and supported me in their own special way over the years: Sengim Karayalçin, Esa Kasmir, Noël van de Ven, Romke Bak, Rory Rinck, Kees Fani, Caspar Krijgsman, Cas Krijgsman (no relation), Lucas Krijgsman (relation to the former), Wolfgang Bubberman, Wang ”Kevin” Su, Abdel Atif, Brennen Bouwmeester, Mickey van Immerseel, Yannick Haverman, Wouter Morssink, Wesley Baartman, Simon de Rooij, Alexander Sterk, Matthias ”DC3” Bakker, Sam van Buuren, Kristof Adriaensen, Matthijs Wisboom, Damian Mercan, Oscar Pater, and of course everyone who helped me with the experiment as well as the kind people of De Kindertelefoon. And finally, a big thank you to my family, who has always been there for me even in times when I felt that I did not deserve it: my sisters Ingeborg Mellonie Braam, Anouk Margot Braam, and my mother Gea ter Haar.

A. A. Braam
Delft, December 2023

Contents

Abstract

De Kindertelefoon is a children's helpline aimed at providing (pre-)adolescents with a person to talk to for a variety of subjects such as bullying, sex, and abuse. These people who talk for De Kidnertelefoon need proper training and guidance. Among the tools that De Kindertelefoon provides is the 5-phase model, a conversational model. Researchers set up a simulation of a virtual child and De Kindertelefoon to help teach the 5-phase model. However, simulation alone was not enough, and we wanted to see if a feedback system could improve the results, with immediate feedback being the focus. Immediate feedback is when the user fills in an answer and the system immediately provides information to strengthen knowledge. When applied to an academic environment, both immediate feedback and student engagement have been proven to be important for completing a task.

Using a literature review, we found that immediate feedback can be directly linked to educational techniques such as self-improvement, self-efficacy, and the Self-Determination Theory. Through a focus group, we also found that constructive immediate feedback is an important pillar of De Kindertelefoon.

The design was achieved by looking at the existing and limited 5-phase model as a graph and trying to find an optimal path through that graph.

After conducting a within-subjects study experiment with 34 participants, the results were inconclusive, with neither condition appearing to be more useful for the group of participants, nor either condition being better at teaching them the 5-phase model. As the results were inconclusive, the data was explored more by looking at it as a between-subject study, which showed that the explanation sheet might perform better for knowledge.

The research shows the possible strength of feedback in a practical manner. From the results, the immediate feedback is neither more nor less resonant than a classical approach to teaching the 5-phase model.

1 Introduction

1.1. Motivation

With psychological problems on the rise according to Higgins [41], there is a growing need for personnel that is both trained and equipped to help and guide people suffering from psychological problems as said by Weiner [82], and Denys [24]. Naturally, this is also true for the younger generation. However, training personnel to help guide this younger generation is not easy. Unlike many help services where explicit help is offered, a child-centred helpline organisation permeates the philosophy of ”we care, we listen”, according to Emmison and Danby [25].

This also rings true in the Netherlands. A children’s helpline named De Kindertelefoon[1] has volunteers who have the task of supporting pre-adolescents between the ages of 8 and 18. Those pre-adolescents can engage with Kindertelefoon volunteers through calling or chatting. Another option is to go on their forums through a web-based application. These conversations are anonymous and in a one-to-one setting. Chat-based conversations would usually last longer than phone calls according to De Kindertelefoon [46]. This could be due to the children feeling a greater sense of engagement when using chat-based, as was found by Sindahl [74]. Sindahl found that chat-based conversation could cause the pre-adolescent to dive into more personal problems. In addition, earlier research by Fukkink and Hermanns [31] showed that some children prefer chatting because it gives them an increased sense of anonymity.

No matter the medium, the topics discussed during the conversation are varied, but common topics include relationships, sexuality, and bullying according to De Kindertelefoon [47]. For handling such sensitive topics with care, De Kindertelefoon trains their volunteers to adhere to their own **5-phase model**. This is described in Beyn [7], which outlines the different phases of the conversation and how De Kindertelefoon counsellor should approach each of those phases. Kindertelefoon on average has a little under 700 volunteers available and in 2021 they trained 224 new volunteers [47]. According to our contact at De Kindertelefoon, this number increased in 2022 to 280 new volunteers. With such a large influx, De Kindertelefoon required a more fluid system.

Chatbots could offer such a system. Chatbots are machine agents that can be used as natural language user interfaces for data and service providers, as stated by Brandtzaeg and Følstad [11]. They go on to say that, most people who used chatbots did so as a means to gain quick and consistent feedback. Chatbots have been used in the fields of both education, as documented by Lidén and Nilros [55], and the training of skills, such as research done by Clarizia et al. [19]. Entenberg et al. [26] even found success in training the parents of children using chatbots. Even though they had a small sample size, many parents noted that it helped train skills on how to use positive attention and praise to stimulate positive behaviours in the child. At the start of 2022, work was done by Grundmann [35] to develop a chat-based virtual agent for De Kindertelefoon. Based upon the Belief-Desire-Intention framework as found in Harbers, Bosch, and Meyer [38], the goal of the virtual agent was to simulate a conversation of a bullied child, to help support the training of possible future helpline counsellors.

[1] https://www.kindertelefoon.nl/

Success was mostly measured through self-efficacy, as it was an efficient and reliable method to assess communication skill training according to Ammentorp et al. [1]. However, the work showed that training with the agent decreased the self-efficacy of the participants.

According to the study, nearly a third of the participants found the feedback provided at the end of a session of little use. Feedback is critical for a person to develop new skills according to McKendree [59]. It is an important step during the learning process, as it allows for self-reflection and professional judgement. In addition, it also helps with the retention of knowledge as found by Kourgiantakis, Sewell, and Bogo [50]. Without feedback, trainees struggle to learn from their mistakes. To that end, we wanted to enhance the previous work done with De Kindertelefoon by creating a feedback system which helped would-be counsellors adhere to the 5-phase model.

1.2. Research Question

The goal of this research is to gain the knowledge needed to create a useful feedback system and integrate it into the existing child virtual agent system. So what are the relevant design factors and constraints that influence the effectiveness and feasibility of the feedback approach? Knowing these factors and constraints allowed us to formulate a design solution, and by answering the research questions defined at the end of this section, we should be able to gain this knowledge. A key example is the knowledge of what makes for useful feedback. For Kindertelefoon we wanted feedback that both increases self-efficacy over time and feedback which helps trainees adhere to the 5-phase model. Then we needed to know the system in which we build and deliver this feedback. As such, we decided upon the following research question:

What feedback-system design is useful and feasible for training children's helpline volunteers?

We split these up into the following sub-questions:

- What feedback-system design is useful and feasible for training children's helpline volunteers?
- How could feedback be integrated in a conversational agent training environment to attain a high perceived usefulness of the system as well as the knowledge of the 5-phase model?
- How useful do trainees find the feedback, and how much knowledge do they gain?

1.3. Research Approach

The research began with a literature study focused on the topics of feedback, feedback systems, and chatbots. This was done to build up a base of knowledge from which the design could be built and see what modern-day technology and techniques have to offer within the field. The literature review then switched to research in children's helplines and education. The former was not only to gain more information, but also to find a helpline perspective outside of De Kindertelefoon's. The latter was to broaden our knowledge for training people in new skills, and to know what else besides feedback is relevant for that task. Finally, research was done on different training systems which emphasized their feedback system, to see how others implemented different theories and techniques, and to learn what each strength and weakness that came with those choices was.

We also worked with De Kindertelefoon and they provided us with experts who helped inform our research (chapter 2). In addition to this, we also set up a focus group where we showed ideas of what a possible feedback system could look like, and by showing several scenarios we promoted discussion between the participants. Summarized and anonymized notes were taken of the participants' thoughts and opinions of the system, which in turn helped shape the final design (chapter 3). The design was turned into a direct feedback system, which was used in an experiment using a within-subject study design (chapter 4). The results found were inconclusive, and we gave our conclusions and recommendations for future works (chapter 5).

2

Foundation

In this chapter we will focus on answering the first research sub-questions:
What feedback-system design is useful and feasible for training children's helpline volunteers?

2.1. Methods

There were three main sources of information used during the research. Literature, and a focus group with the experts of De Kindertelefoon. To try and avoid biases, methodological triangulation was used to have multiple methods of data collection. Bekhet and Zauszniewski [6] state that methodological triangulation can be useful for providing the conformation of findings, increased validity, and a better understanding of the studied phenomenon/subject. In accordance with Thurmond [77] and Casey and Murphy [16], a within-method was used as both the focus group and literary research could be seen as qualitative methods.

2.1.1. Literature Study

For the literary research, Google Scholar and WorldCat Discovery were used as the primary search machines, and additional literature was also provided by the supervisors. Important findings were methods and techniques such as the Self-Determination Theory and the aforementioned methodological triangulation, as well as literature on virtual agents and coaches. Search terms used included ”self-reflection,” ”failure in learning,” ”immediate corrective feedback,” ”educational techniques,” ”chatbots for training,” ”scaffolding in skill learning,” ”helpline training,” and ”self-efficacy in education.” Papers were more readily included if they included subjects on training or working with children specifically. Papers that discussed some of the applied techniques such as scaffolding would specialize too far into a particular field, e.g. medicine, as to no longer be applicable for training children's helpline counselors, and were rejected for this study as a consequence. There were roughly 3 themes/subjects on which we were focused. Firstly, we investigated tutoring and educational methods, which were necessary to understand how people teach and learn new skills. Furthermore, we looked into chatbot systems in medical and educational contexts, to see what the current state of technology is and how it has been implemented and received so far. Finally, we looked into the dissection of what feedback is and how it helps in acquiring new skills. These points helped to set boundaries and expectations from which certain design factors/concerns were created.

2.1.2. Focus group with the Experts of De Kindertelefoon

The design factors/concerns were also derived with the help of the experts of De Kindertelefoon. A research associate and psychology student also provided us with De Kindertelefoon's training materials and helped set up the focus group. This focus group was approved by the TU Delft Human Ethics Research Committee (HREC reference number: 1289)

During the focus groups, we showed different scenarios of what a prototype of a possible feedback system could look like and made summarized notes of the opinions and ideas of the focus group. In turn, these opinions and ideas helped inform the design factors.

Participants

For the focus group we met up with three senior members of the organization, all were experienced volunteers and two of them also had experience tutoring new trainees. They were all grouped together for a single focus group session. Also present during the focus group was the aforementioned psychology student contacted by De Kindertelefoon who helped notify the participants' opinions and ideas.

Materials & Procedure

The meeting was held online on Microsoft Teams due to the volunteers living in different parts of the country. As a whole, the focus group took about an hour to complete. The group was presented with 2 scenarios that are useful in the design process as they show the consequences and trade-offs of different designs [15]. The scenarios show a narrative that can help the participants understand the situations and contexts for the designs better [67].

To ensure that each participant took part in the discussion they were asked to individually fill out an online form after each scenario. The forms consisted of two questions each. The first question asked them to rate their agreement with a statement related to each scenario using a 4-point Likert scale. The second question was for them to note down why they agreed with certain topics. The former was to ensure that participants would not suddenly change their answer after hearing one of the other participants speak their thoughts and to make them pick between the two systems instead of trying to find some middle-ground. The latter question was to help the participants remember why they picked an option. Notes were taken by two people who later compared and checked them off each other, to minimise any biases. The ideas and opinions of the summarized notes were used to set the design factors of the feedback system and to help set up the expectations of Kindertelefoon for what the system could do.

The 2 scenarios were based on the literary research shown throughout this section and refined through meetings with supervisors and colleagues. One scenario focuses on the difficulty of the conversation and how much support the user should get from the system, with the goal being that it would show what kind of skills the focus group values in their volunteers and how to measure the improvement of the skills. The other scenario focuses on the feedback and was used to gain insight into how the focus group provides feedback and what they find important about the feedback they give. Each scenario showed a different aspect of a possible feedback system and two ways of implementing it. These scenarios were then discussed by the focus group which was observed. In an open discussion, every participant was free to explain what they thought about the scenario and to express their concerns about each scenario. The discussion's goal was not only to gain opinions and prompt discussion regarding the design of functionalities of the system but also to learn what values they cared about in a feedback system. Figuring out what motivated the experts' answers is more important than the answers themselves, as it gives us more to work with for the design factors. The complete collection of slides used for the focus group is in Appendix A.

2.2. What is good feedback?

Before discussing the results of the focus group, it is important to explain the building blocks of the different types of feedback that were discussed with the focus group. The basic idea as described by S. Grundmann [35], is outlined in Figure 2.1. A trainee gets the theory explained to them, which consists of each step of the 5-phase model, and the different tactics and scenarios that go with each step. What followed is a role-play session with the virtual agent in which the trainee tries to apply the theory into practice as the trainees try to guide the simulation of the bullied virtual child through the 5-phase model. At the end of the simulation, feedback was given to the trainee, a briefing containing a transcript of the conversation as well as the so-called beliefs of the virtual agent. For this research project, the idea is to also give ”live feedback” during the role-play session. This is the bottom part of the image where the role-play determines what feedback the system gives the trainee. This feedback would be more elaborate and focused on the trainee's performance rather than the cognition of the virtual child. A large part of the design is about finding out what sort of feedback would work best for De Kindertelefoon and how it should be implemented.

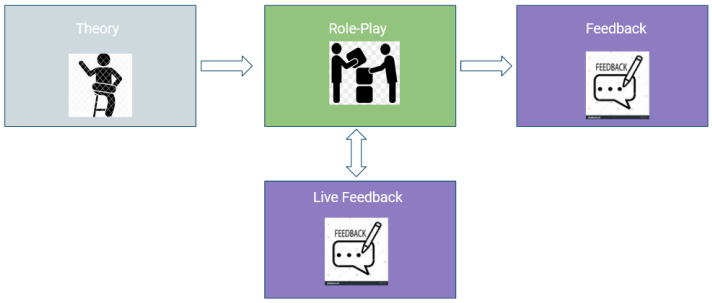


Figure 2.1: New system with at he bottom the added ”live feedback”

2.3. Immediate and delayed Feedback

Two possible options for feedback are immediate feedback and delayed feedback [51] [63] [18]. Immediate feedback is when the user fills in an answer and the system immediately lets them know whether their answer was right or wrong. Delayed feedback is the reverse of this where the user gets the feedback after they have completed the training session or a certain amount of challenges. This brings us to Epstein et al. [27], who conducted research where the authors immediately showed a user whether their answer was right or wrong. Their research showed that immediate feedback resulted in increased retention of information and caused users to be more accepting of feedback in general. Something similar was found by Pashler et al. [64] who noted that immediate feedback after an incorrect response improved retention. On the other hand, Metcalfe, Kornell, and Finn [61] say that delayed feedback when a user gives a correct answer can help the trainee retain the information better. Similarly, research from Van der Kleij et al. [79] states that immediate feedback is better for memorizing facts, while delayed feedback is better for applying or transferring knowledge.

Through the focus group, we found that some preferred immediate feedback because they believed that De Kindertelefoon taught through ”redo’s.” The members of the focus group said that allowing Rob to immediately apply the correct method to a mistake he made and experience a sense of success from it was a key feature of learning. Scheeler and Lee [73] argued that corrective immediate feedback was used to positive effect on would-be teachers in their paper, teachers who received the immediate feedback were more successful in finishing the experiment’s trial than those who were not. Other participants of the focus group voiced interest in the delayed feedback because they were concerned that Rob would not be able to learn from his mistakes. These members said that the best way to learn the importance of the 5-phase model was to experience the consequences of not applying it properly. While that sounds logical, Eskreis-Winkler and Fishbach [28] argue that people learn less from failure as it causes them to tune out and take up less information. Eskreis-Winkler and Fishbach go on to say that small amounts of corrective feedback can work fine, but that any educator should be wary of the build-up of negativity that corrective feedback can cause.

The focus group comments also seemed to indicate that they thought of feedback only as a corrective measure, but did not consider that it has motivational qualities as well. When positive reinforcement as feedback was brought up, the focus group felt that positive direct feedback would work well. But this was only the case if the feedback was not a constant presence, as the focus group feared that too much feedback, of any kind, would distract from the flow of the conversation and disrupt the 5-phase model. The reasoning the focus group gave was that the flow was important to the trainees as they had to perform the different phases correctly one after another. If a previous phase went poorly, the focus group could not imagine a later phase going well either. To add to this, Scheeler and Lee [73] also gave feedback in low frequency to their preservice teachers, though still immediate.

Naturally, a hybrid of immediate and delayed feedback could be potent, as stated by Robinson et al. [69], where feedback depended on the state of the user. Both types of feedback were used to create a stronger overall product. In a system developed by Courgeon et al. [21], another hybrid where feedback was provided both immediately and after examination, the authors learned that it is necessary to keep the feedback system neutral during the examination, but moved away from the neutrality after the examination was over. The neutrality, meaning that the tone and the use of vocabulary are neither

aggressive nor patronizing, during the examination was important, as otherwise, the feedback could potentially damage the trainees' confidence, while less neutrality and more energetic feedback during the delayed feedback helped with retention.

During the focus group, a similar technique was brought up, where they would be less confrontational with their feedback during examination in order not to disrupt the flow of the trainee, but would give more detailed feedback between different phases.

2.4. How to measure people's improvement?

Whether gained through direct or indirect feedback, people's mastery of the 5-phase model is important to consider. Which makes it necessary to measure how much their mastery improves.

2.4.1. Self-efficacy

The previous work, Grundmann [35] focused on the virtual child and tried to measure using self-efficacy. In counseling training, self-efficacy refers to a counselor's beliefs about their capabilities to effectively counsel a client in the near future as defined by Larson et al. [54]. However, after interacting with the virtual agent [35], self-efficacy in the trainees went down for some. There are a few explanations for this occurrence, one of which is that the participants lacked the necessary real-life experiences to properly estimate their self-efficacy. Something similar was mentioned by one of the participants of the focus group, who was of the opinion that people often are not great at self-estimation. Furthermore, Grundmann [35] did state that they saw a significant decrease in self-efficacy after the virtual agent training, which was not found with text-based training.

Studies have shown that feedback can be beneficial in both increasing self-efficacy according to Karl, O'Leary-Kelly, and Martocchio [49], as well as moderating it according to Beattie et al. [5]. As such, it is reasonable to state that feedback can lead to better performance and higher self-efficacy. Furthermore, Billings [9] also found a positive link between higher self-efficacy and improved performance of their participants. Feedback would result in people performing better more quickly and it would appear to be crucial to get feedback at the start of training.

De Kindertelefoon also gave an important pointer on a philosophy that they incorporate from the start of training. De Kindertelefoon has a philosophy that someone has ownership over their learning process. As such, their feedback usually accommodates this. Rather than just correcting someone, feedback should push counselors-in-training in the right direction and allow them to reach the answer on their own. As a result, some also adhere to the idea that counselors-in-training should be allowed to try and experiment, and undoubtedly fail. This could have merit, as learning from failure is a method used in a variety of fields such as work by Anderson et al. [2], and Sawyer [72]. However, this was hardly a popular opinion, as others mentioned that early success was a good motivator and could provide energy and determination to trainees later in the learning process. According to one participant, this was because they wanted the counselors-in-training to ”own their own path of learning”.

2.4.2. User agency

Agency is directly related to this idea of owning your path of learning and learning through your own mistakes. Agency is the ability to act independently. In the process of learning, Biesta & Tedder [8] describe it as the level of quality of the participants. Agency is very key at De Kindertelefoon as an institution. When the participants of the focus group were pressed on the matter of ”owning your path of learning”, they said that there was a desire to train the would-be counselors in the same manner as the pre-adolescents are spurred into action. An important part of the 5-phase model is that the counselor should not come up with solutions for the child. Instead, the counselor is there to help guide the child toward taking action. In other words, De Kindertelefoon wants to promote the child's agency.

Closely related to this agency is the Self-Determination Theory (SDT). According to Weiner [83], SDT is a popular approach to studying human motivation. From the very beginning, SDT was founded on research looking at what sort of effects intrinsic and extrinsic rewards have on a person's motivation. Ryan and Deci [70] found that people who received positive feedback on their competence could have increased levels of intrinsic motivation. Furthermore, Ryan and Deci [70] also state that SDT specifies

three basic psychological needs. Satisfaction of these three psychological needs has been described by Deci and Ryan [22, 78] as necessary for maintaining both intrinsic motivation and extrinsic motivations. The first need, and the need we will focus on, is the need for autonomy, which is when a person feels that they have a choice and that they willingly can perform said choice from their own volition. In other words, a person feels that they are the origin of their own actions.

This is similar to what was found by McKevitt et al. [60], who found that in order to engage students, tutors should allow students to assess their own work and thus give them a sense of agency. The feedback should not just be about right and wrong, but also about how the trainees can reach the right answer and how could they have seen what the right answer should have been. Cohen et al. [20] describe the former as cognitive feedback, which states how to perform a task, and the latter as feed-forward feedback, which helps the user anticipate different decision options.

When an experienced counselor of De Kindertelefoon was asked to comment on their experiences they mentioned that whenever they would ”get stuck” on a subject during role-play, their trainer would pause the session to give feedback, let the counselor-in-training think about it, and then resume the session. This feedback was described as ”making you think about a different approach”. This means that instead of flat out giving the answer, the trainer chose to let the counselor-in-training come up with their solution. Thus promoting their agency and, in their own words, ”good for raising their confidence”. Our design could have something similar, where feedback is given after a trainee has not responded for a while, to get them ”unstuck.”

However, this giving feedback when someone does not respond for a while must not be overdone. Hays et al. [39] found that feedback at times is not beneficial, and can even be detrimental. Hays et al. found that allowing the users to skip certain feedback allowed for a greater self-reflection on the users' part as well as a more critical view of their own capabilities. Robinson et al. [69] also found that users with a high self-efficacy preferred to skip feedback as it would merely interfere with their ability to complete the task. The design could thus give less feedback to people that are performing well, which in our case would be to smoothly go through the conversation, while correctly applying the 5-phase model

2.4.3. Levels of skill and knowledge

Besides self-efficacy and agency, there are also other factors to consider during training. Namely, knowledge and skill. Knowledge is about the different types of techniques that can be applied at each phase of the 5-phase model and the skill to properly apply them. As the training progresses, a trainee's skill and knowledge on the subject should increase. Naturally, this also affects the feedback.

At the start of their training, trainees presumably have little to no knowledge of the 5-phase model. As such, their knowledge has to be built up first before they can use the knowledge for more complex techniques. For example, first, the new trainees need to know how to ask questions with which they can gain information from the child, and only then can new trainees learn how to use this information to formulate follow-up questions. This division and hierarchy of things to be learned are, in a sense, a learning method called scaffolding. As explained by Gonulal et al. [34], scaffolding is a technique that has gained popularity in several education-related fields. The original idea was focused on children, where a parental figure made timely interventions for children performing problem-solving tasks, letting the children figure things out for themselves for the most part. The expert of De Kindertelefoon experienced this as well. After the experts completed the training sessions, they would then move on to actual conversations with children. During these conversations, a mentor would sit close by to monitor and intervene if necessary. Gonulal and Loewen [34] goes on to mention skills that scaffolding techniques could help improve, such as maintaining goal orientation, controlling frustration, and modeling solutions to the task.

As knowledge and skill improve over time, training needs to be adjusted to account for this increase. A discussion arose in the focus group when asked how knowledge and skill improvement should be measured. Some wanted Rob to gain the ability to self-reflect and gauge his own level of skill, while others preferred to have a tool to measure it. Those who favored a tool of measurement said that it was because, from their personal experience, people were inherently not good at gauging themselves. These members of the focus group went on to say that the ability to self-reflect was important and that

they did not necessarily disagree with the other members, but rather that they believed that it had to be taught and that it was not something inherent to most people, also that teaching people to self-reflect on top of handling the 5-phase model was outside the scope of their capabilities. Von Wright [80] agrees with this point, saying that self-reflection is a skill hard to come by but that methods such as reciprocal teaching can be applied to help teach self-reflection. In a more recent study, Gün [36] also showed that training self-reflection through feedback from peers and watching recorded videos of oneself amongst other elements can help someone train self-reflection. When asked why self-reflection was so important, the group agreed that self-reflection as a skill was needed, as without the ability to reflect on your own decisions and processes you cannot help others with theirs. Hilden and Tikkamäki [42] state that the consensus at their time was that reflection was at the core of learning and professional growth, in agreement with the statements of the focus group. In addition, a paper written by Brownhill [13] states that as a skill, self-reflection has only become more important over the years.

2.5. Main outcomes

So what makes good feedback? O'Donovan et al. [62] found that what makes feedback "good" is highly contextual. In the case of De Kindertelefoon, the focus group experts argued that they want to teach their trainees in the same manner as the pre-adolescents are helped, by making the trainees feel that they own the training regime. This means that the system should promote the users' agency and self-efficacy, which is strengthened by the Self-Determination Theory, with them being the psychological need for autonomy and competence respectively. Ryan and Deci [70] goes on to state that as a whole, SDT provides a detailed framework for understanding human agency, and we want to promote that agency using feedback.

It has also been established that feedback can be both immediate and delayed. Immediate feedback can be in the beginning when the trainees still have much to learn and when the amount of errors made is at its highest. This immediate feedback could be given with a neutral disposition. Meaning that it should not be given in an accusatory manner, but rather as a matter-of-fact. This is because De Kindertelefoon wants to keep their would-be counselors motivated and not overwhelm them with negative feedback. De Kindertelefoon wants to promote positive feedback often believing it to be the most important one for information retention, as some of the found literature reflects this as well, such as Metcalfe, Kornell, and Finn [61], and Van der Kleij, Feskens, and Eggen [79]. Feedback should also be built up. Because, as the trainee moves along further in the training program, their skills and knowledge mature and the feedback should reflect this.

Table 2.2 contains an overview of the found design factors/concerns through all the different methods, as well as the reasoning as to why they are important. Also given is an idea for the design which will be expanded upon in the next chapter.

Design Factor/Concern	Why	Design Idea
Immediate feedback, both positive and corrective, is important.	Immediate feedback is important to learn to overcome errors. Furthermore, moments of success are key at keeping people motivated throughout training.	If a trainee makes a mistake, the system gives them the information needed to find the correct answer and then let them redo it. Corrective feedback can be given, but done so with tact.
Immediate feedback needs to be given with a neutral disposition.	We do not wish to disrupt the flow of the conversation or make the users worry too much about the feedback.	The feedback is not too friendly or aggressive, but delivered in a matter-of-fact disposition. Look to De Kindertelefoon for correct use of language.
Both types of feedback should not be given in an overwhelming amount.	Too much feedback would disrupt the flow of the conversation and could overwhelm the trainee.	Immediate feedback can be given when a trainee gives incorrect input, but immediate positive feedback is only given at set intervals. Delayed feedback is succinct and can be given at the end of each phase.
The system could try to promote the user's agency, self-efficacy and autonomy.	Agency helps both with motivation and retention.	Feedback could give contextual clues to the answer, instead of the answer outright.
Skills improve over time and that improvement could be tracked.	People lack the ability to self-reflect and they lack the means to train people to gain that skill.	As the model is closely related to the 5-phase model, it can be worthwhile to focus on the trainee's knowledge and level of application of said model.

Figure 2.2: Design factors with possible design ideas for the future

3

Design

In this chapter we will focus on answering the second sub-question:

How could feedback be integrated in a conversational agent training environment to attain a high perceived usefulness of the system as well as the knowledge on the 5-phase model?

The design factors of the system should thus reinforce certain measures. It would reinforce the perceived usefulness and knowledge of the 5-phase model. The design factors also need to properly integrate with the existing virtual agent. Grundmann's [35] chatbot is a BDI-based virtual agent. BDI stands for Beliefs, Desires, and Intentions. *Beliefs* are what the virtual agent thinks is true about the world, *Desires* is what the virtual agent wants to change about the world and an *Intention* is the desire the agent acts upon. By interacting with the chatbot, people alter the beliefs of the chatbot, which in turn changes their desires, eventually leading to the conversation's end.

The main idea is to steer people towards interactions, that alter these beliefs and desires, such that the desired end of conversation is reached. The beliefs and desires become a "path" that the user can traverse. Every input given by the user is a step, and there is a set order of steps which is the "best" path. If the user follows the path, then the system responds by giving the user immediate positive feedback. If a user would stray from the path, then that becomes an opportunity for immediate corrective feedback. Both these types of feedback will influence the user interaction once again, creating a loop as shown in Figure 3.1.

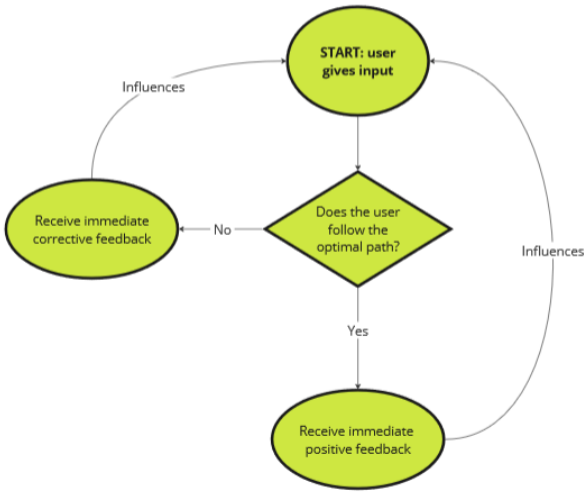


Figure 3.1: High-level overview of the system and user interactions.

In this chapter, the design and implementation of the immediate feedback will be discussed in greater detail. Parts that will be discussed, are how the system is able to detect when it should give feedback, what information the feedback should contain, and how the feedback should visually be brought to the users.

3.1. From the BDI-model to the path

At the start of the design phase, it was clear that the existing BDI-model would stand at the center of whatever system would be added on top of it. The BDI's were modeled to the 5-phase model and a user's input can change the values of these beliefs. These desires were devised as a blueprint for how the conversation should go. Lilobot starts off by wanting to talk about their problem, otherwise, they would not seek out contact with De Kindertelefoon in the first place. As the counselor in training progresses through the 5-phase model, Lilobot will come to believe that they can trust the counselor. At this point, Lilobot asks for the counselor to call the school to remove the bullies for them. Counselors cannot interfere in such a manner according to De Kindertelefoon policy, leaving the counselor's only correct option to reject Lilobot's proposition, causing a decrease in trust. The counselor needs to regain Lilobot's trust and motivate them by opting to work together on a new solution, which will eventually lead to the child wanting to talk to their teacher and leaving the conversation. As such, all of these desires need to be hit for the conversation to have followed the 5-phase model. Lilobot switches between desires based on the current values of the beliefs, which get altered by user input. Lilobot keeps track of each belief as a value between 0 and 1. When certain belief values reach certain thresholds, the Intention (the active Desire) changes, and with it the actions that the virtual child takes. For example, when Lilobot's beliefs "I have told my story" and "I can trust De Kindertelefoon" are high enough, Lilobot will switch from desire D1 to desire D3 as shown in Figure 3.2. Also in Figure 3.2, you can see the 5 different desires as set up by Grundmann [35], and the order (D1-D3-D5-D4-D2) in which they need to be traversed.

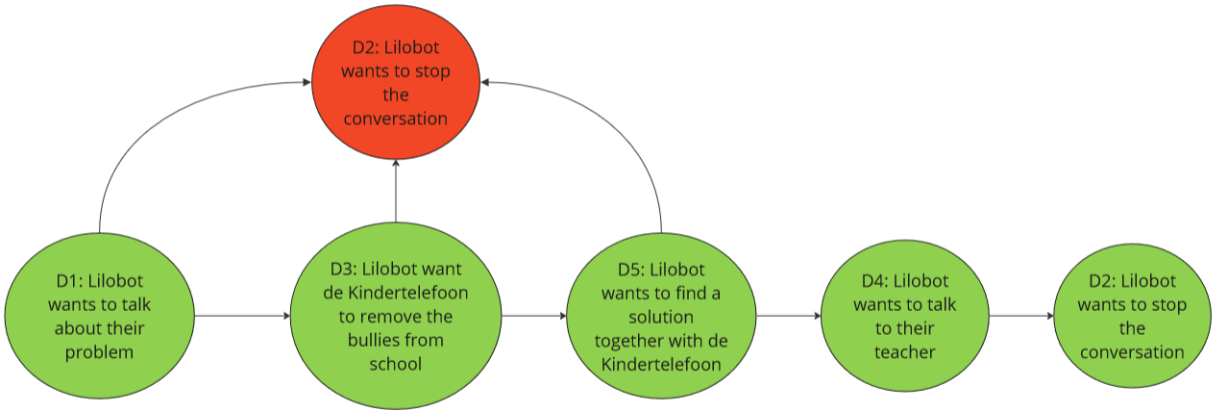


Figure 3.2: Correct order of the desires, with possible side paths. Green denotes the correct path, while red shows an undesirable end.

As can be seen, the only possible way to skip into a wrong phase is D2, Lilobot wants to end the conversation. Reaching D2 is only desirable when D4, Lilobot wants to talk to their teacher, has already been reached prior. As it currently stands, this is the optimal situation in which Lilobot establishes enough confidence such that it can solve the problem itself. Going into D2 from any other phase means that Lilobot either believes the counselor to be threatening, or that the counselor is unable to help them with their problem. For the correct sequence of desires to occur, a particular order of belief values has to be met, to smoothly navigate the conversation. Thus, solving this would be akin to a graph solution known as ordered Traveling Salesman Problem (oTSP) [52], a well-explored algorithm [53], which is still important to this day [48]. The graph here would be all the Desires the counselor has to move through, starting in Desire D1 and ending in Desire D2 after visiting all of the other vertices

(the other Desires). The edges between the vertices can then be seen as the combination of user inputs, which alter the Beliefs of Lilobot in such a manner that Lilobot goes from one desire to the other. Another option was a variation of this idea, where an adaptable algorithm would calculate the shortest path every time a person made an input. However, due to the limited amount of fields the conversation can go in, it was decided to go with the handcrafted path.

3.2. General overview of the path

An overview of how people interact with the handcrafted path can be seen in Figure 3.4.

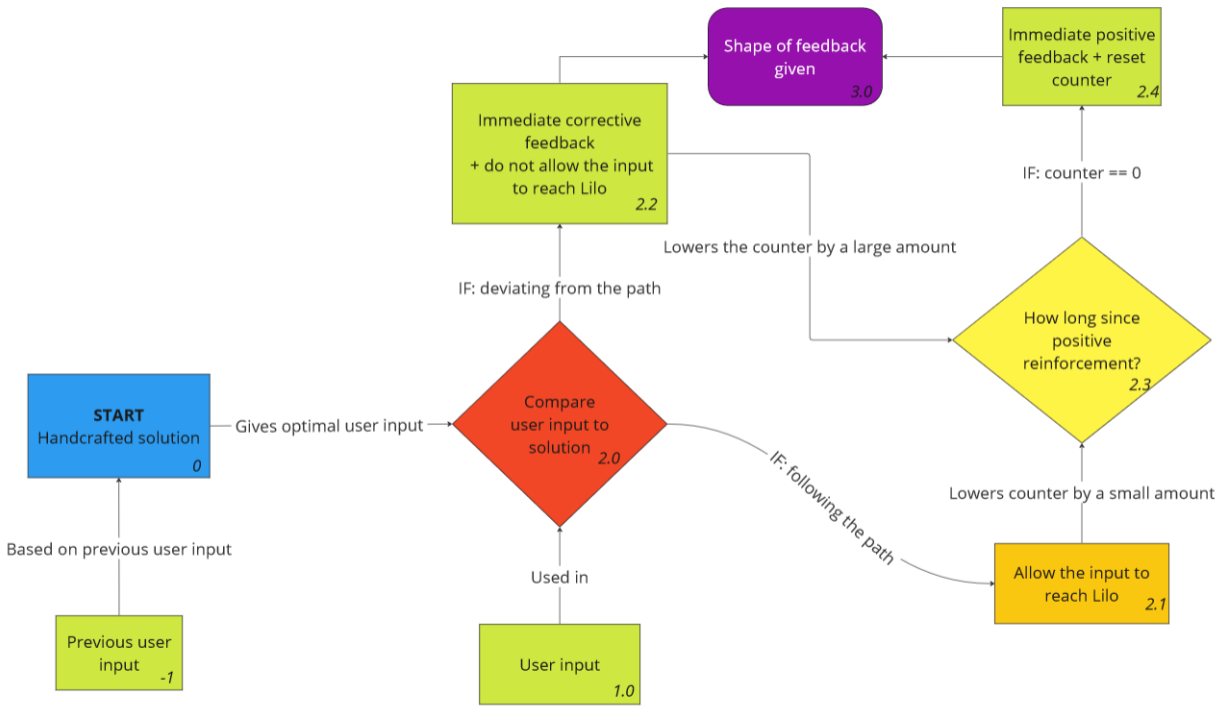


Figure 3.3: High-level overview of how the feedback system will determine the correctness of the user input system

At any point during the interaction, the system looks at the previous inputs of the user (box -1). Based on those inputs, it is determined where the current conversation is within a handcrafted solution, from which the system can see what the next best step would be in that handcrafted solution (box 0). This best next step is compared to what the user inputs into the system, see box 1.0. This user input is used in the comparison to the handcrafted solution. This comparison happens in box 2.0, where it is checked whether what the user inputted is the same as what the handcrafted solution says is best. This can result in a few possible outcomes as shown in boxes 2.1 and 2.2. In box 2.1, the user input and the step on the handcrafted solution match, meaning that the user follows the path, which means that the input should be allowed to ”reach” Lilobot. Reaching here means that the input should be allowed to change the Desire-related beliefs from the BDI-model, as this, according to the handcrafted solution, would be a good change in the internalised belief of the chatbot. If box 2.2 applies, the user input should be stopped from reaching Lilobot in order not to alter the Desire-related belief values. This would allow the user to do the interaction again, but now with the corrective feedback, allowing for a moment of reflection and learning. Both box 2.1 and 2.2 influence box 2.3, which is a more technical aspect of the system. Internally, the system kept track of a counter. When that counter reached 0, positive feedback regarding what the user did would be given, as can be seen in box 2.4. This internal counter gets lowered whenever the system reaches box 2.1 and box 2.2. Box 2.1 lowers it by a small amount, as not every correct input warrants positive feedback from the system. Box 2.2 however, lowers the counter considerably more. This was done to users feel better that they succeeded after several failed attempts. Together, boxes 2.3 and 2.4 shape the feedback given to the user (box 3.0), which signals the end of the interaction before the cycle begins anew.

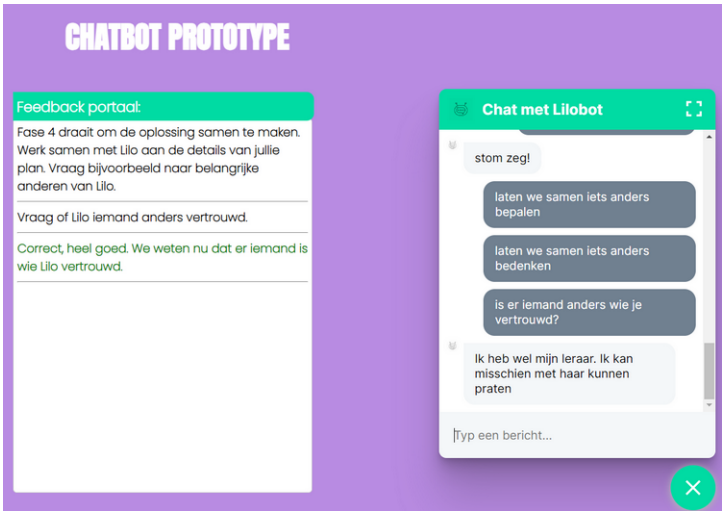


Figure 3.4: Box 2.4 executed. After a few erroneous inputs, the counter reaches 0 and the system gives positive feedback to the user.

3.3. Shape of immediate feedback

When talking about the shape of the immediate feedback, besides its mere contents it is also about how the feedback is given, and how to integrate the feedback with the existing system.

3.3.1. The theory of corrective feedback

Corrective feedback is potent when applied directly after a trainee makes an erroneous input according to Epstein, Epstein, and Brosvic [27]. After it is made clear to a trainee that their input was erroneous, they should receive feedback and be allowed to interact with the bot. Erroneous input which could cause the situation to go amiss should be prevented from altering the BDI-model. This preserves the state as it was prior to the erroneous input, allowing the user to re-do the interaction until a correct decision has been made. The user is steered towards this correct decision, through the corrective feedback which is determined by the state the conversation is currently in. To add to this, this corrective feedback should be of the type which does not give the answer outright, but rather gives the trainee the information needed to find the answer themselves, as well as telling them contextual clues from which they could have found the answer.

However, as most of the people who partake in using the system are new to both it and the 5-phase model, it seemed likely that people could easily make mistakes, get stuck, and lose motivation. In order to preserve motivation and give people new to the 5-phase model more of a helping hand, the system gives more direct hints after several failed attempts to progress the conversation, thus steering newcomers more easily toward a proper solution.

3.3.2. The theory of positive feedback

On the other hand, while positive feedback is also important, too much of it can negatively affect the experience, as explained in Chapter 2. To control this, the system only gives positive feedback if the trainee has not received any such feedback in a while. This is tracked within the system with an innate timer, which is lowered by values depending on which type of feedback is received. This was done to give trainees who are struggling a moment of success, by pushing them towards finding a correct solution on their own. A success-breeds-success theory as found by Salanova, Martínez, and Llorens [71] and Iso-Ahola and Dotson [45].

3.3.3. Design of the visuals

Whether it is corrective or positive feedback, it should be displayed in a visually appropriate manner. The design of the interface impacts how the user perceives and receives the feedback [43]. As such, putting the feedback at the center of the screen ensured that people would not miss the information provided to them. The feedback was shown in a small text box of similar size to the chat screen, but

visually distinct from the chat screen due to colouring. The feedback was also made to fade into the log rather than pop in instantaneously, in order to be less stressful.

In the final prototype, the user interface is comprised of 3 major parts. On the left-hand side, the participants can find a short explanation of the various do's and don'ts of the program. This includes giving the chatbot some time to react if they appear slow and keeping the sentences simple. In the middle is the feedback portal which shows the immediate feedback. Upon receiving feedback, the top part will change colour for a moment, notifying the participant that feedback has been given. If the feedback is positive, i.e. the participant input is correct, the text will be displayed in green. While if it is corrective it will be given a neutral black colour. As the underlying chatbot is built very much on the participant asking a question and the bot responding, any feedback shown will only be part of one ask-response part of the interaction, in order to remove clutter as the entire conversation moves on. Finally, on the right, the participant has a chat window that they can use to interact with the chatbot.

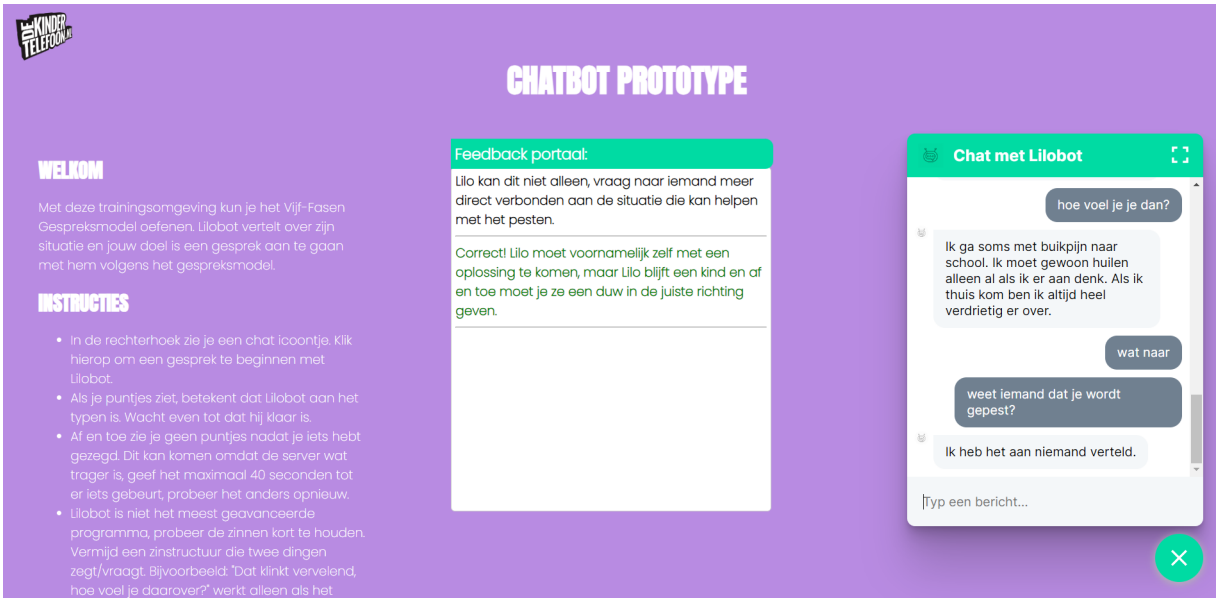


Figure 3.5: From left to right: Sidebar text explaining the prototype, the feedback portal, and the chat window. After the user asks Lilo if anyone is aware that they are being bullied, the feedback reads: "Correct! Lilo should come up with the solution themselves, but they remain a child, and they might need the occasional nod in the right direction.

4

Evaluation

This chapter will answer the sub-question:

How useful do trainees find the feedback, and how much knowledge do they gain?

To help answer the research question, several questionnaires were created for the experiment, asking the users to fill in both the perceived usefulness of the system, as well as testing them on the knowledge that they acquired over the experiment. The search for a variable that can be measured for both terms led to Igbaria and Iivari [44], who developed a Technology Acceptance Model (TAM). Igbaria and Iivari created a model focused on self-efficacy which mediated between the two variables *perceived usefulness* and *perceived ease of use* as defined by Marangunić and Granić [56]. Among their findings, they discovered that perceived usefulness strongly influenced usage. Congruent to that are the findings of Bandura [3, 4], who argues that both perceived usefulness and perceived ease of use have a direct effect on motivation, and according to Burgers et al. [14] positive feedback boosts motivation through certain needs. From which comes our first hypothesis:

H1: People who receive immediate feedback during a conversation find the feedback more useful than having an explanation sheet next to them during said conversation.

To have a more quantifiable measure of how effective the system was, knowledge gained on the 5-phase model was chosen. This seemed fitting, as the 5-phase model is the driving force for completing a conversation with the virtual child successfully. The measuring of the participants' knowledge of the 5-phase model was also useful for the trainers as discussed by the focus group, giving us our second hypothesis:

H2: People who receive immediate feedback during a conversation have more knowledge of the 5-phase conversational model than those who receive an explanation sheet of the 5-phase model.

To evaluate the developed prototype, an experiment was set up to investigate the users' opinions of the feedback system. The system's implementation is first sketched out before the experiment setup is established, and the chapter concludes with the experiment results.

4.1. Methods

The experiment to evaluate the usefulness of the feedback was run in July 2023. Before running the experiment, we registered the design of this study with the Open Science Framework (OSF) [1]. Furthermore, the study design was approved by the TU Delft Human Research Ethics Committee (HREC reference number: 1289).

4.1.1. Study Design

The study was a two-group within-subjects study design. During the conversations, the participants received two conditions: either feedback from the system or an explanation sheet. The condition with feedback is shown as text on the user interface. It is about the 5-phase model and what the user can do during each phase, based on where they are in the conversation. The other condition is an explanation sheet, which is an instruction containing the main goals of the 5-phase model. The participants were

[1] https://osf.io/

split into two equal groups. The only difference between the groups is the order of the conditions with which they went through the experiment. One group started the experiment with immediate feedback, filled in the questionnaire when they were done, and then did a second experiment with the explanation sheet and filled in a questionnaire after that. The other group did it with the conditions reversed.

4.1.2. Materials

Before the participants had any interaction with the chatbot, they were required to watch a 3-minute long video explaining the different phases of the 5-phase model. The video was in English as it was developed by another researcher working on a similar topic [2].

The questionnaires were designed using the guidelines as proposed by Brinkman [12]. The questionnaires and informed consent forms were hosted on Qualtrics. They can both be found in full within Appendix C.

Finally, for one condition the participants had an explanation sheet, containing information on the different phases of the 5-phase model, as well as examples, next to them. Due to agreements with De Kindertelefoon, this sheet cannot be shared publicly.

Prototype

The prototype design was implemented as described in the previous design chapter. The prototype is an assortment of several programming instruments. These included: the HTML/JavaScript/CSS trifecta for the webpage aspect where the user interacts with the chatbox and receives the feedback, Python for the RASA integration, and JAVA as the main workhorse taking care of the BDI logic and how the user interactions change those BDI-values. The ones extended for this project were the webpage and the Java code.

The chatbot was running on Rasa version 2.8.1. and ran on the researcher's personal laptop. The code can be found online[3].

4.1.3. Measures

Perceived usefulness.

This was a measure of how the participants perceived the usefulness of the two conditions. It was measured using the questions as provided by the Intrinsic Motivation Inventory (IMI) [57] [4]. The IMI is described as a measurement device intended to assess participants' subjective experiences related to activities in experiments. In total, there are six different sub-scales for assessing different subjective values. For perceived usefulness, the tool's value/usefulness questions were used. IMI uses a 7-point Likert scale to gauge how useful the participants found the advice material. These range from: (1) not true at all, (4) somewhat true, and (7) very true. The IMI was kept intact to preserve the integrity of the system. The only change being a clarification of the condition being tested. In the same vein, despite the experiment being conducted in Dutch the questions of the IMI were kept in English instead of being translated to adhere to the IMI's integrity.

Knowledge regarding the 5-phase model

This measure was based on the knowledge of the 5-phase model. It was measured using a combination of open-ended and multiple-choice questions. Both of these were based on training materials provided by De Kindertelefoon. The multiple-choice questions were each worth 1 point, while the open-ended questions were worth a maximum of 2 or 3 points, depending on how complex we deemed the questions. In order to not let the open-ended questions weigh more than the others, the results of those questions were normalized. The maximum score a participant could reach was 11. Bloom's Taxonomy, as described by Forehand et al. [29], was used to define the questions to cover the layers of remembering, understanding, applying, and analyzing. The levels, as defined by Forehand [29] were used as inspiration for the questions that the system could provide. An overview of the different questions and their levels can be seen in Table 4.1.

[2] link provided by the creator: https://www.youtube.com/watch?v=t6OJ5RYXXIk

[3] $https://gitlab.ewi.tudelft.nl/in5000/ii/child_helpline_feedbacksystem_training_aryton$

[4] https://selfdeterminationtheory.org/intrinsic-motivation-inventory/

Question	Bloom level
What can a counselor NOT do in phase 2?	Understanding
What is NOT a possible action in phase 4?	Understanding
What sentence can you say to the child in phase 5?	Understanding
What should a counselor NOT do in phase 1?	Understanding
What should a counselor NOT do in phase 2?	Understanding
What is NOT a part of phase 3?	Understanding
What is the purpose of phase 1?	Remembering
You have just given the child a warm welcome, what happens directly after that?	Remembering
You have just asked the child what their wish is, what happened directy prior to this?	Remembering
What is the purpose of phase 3?	Remembering
What is the purpose of phase 2?	Remembering
What is the purpose of phase 4?	Remembering
You are now finishing the conversation, what happened directly prior to this?	Remembering
You have just finished the 2nd phase. At the start of the 3rd phase, what sentence could you say to the child?	Applying
You and the child have just made an agreement that the child will talk to their parents, and thus finished the 4th phase. What is the most logical to say now?	Applying
The child asks if you can call her parents to solve the problem for her. How would you react, and why? *(open-ended)*	Analyzing
The child has told you that she is being bullied, before you can proceed to the 3rd phase you first have to gather more information. What sort of question could you ask the child. Give a maximum of 3 answers. *(open-ended)*	Analyzing

Figure 4.1: The questions (translated to English) with their respective Bloom levels

For the questions regarding knowledge, Haladyna [37] and Brame [10] were also used for creating the multiple choice questions. These included guidelines on how to write questions clearly, not allow the answer of one question to inform another question, and on how to avoid having one answer stand out as either the correct one or a red herring. Two versions of the questionnaire were created to decrease biases. For these two versions, the existing questions were mirrored and inverted to help create more questions, and then all the questions were split between the two questionnaires. The questions were posed in Dutch to adhere to the examples found in the chatbot (which was made for the Dutch language).

The open-ended questions were tested using double coding to check their reliability. The first coder was the main researcher, and the second coder was a mechanical engineering master student. Together, the first and second coder agreed on a grading scheme for the questions and then each went over the answers provided by the participants and graded them individually. Then, the grades were compared and discussed, creating a final grading for every participant. As the outside party had no knowledge of the 5-phase model prior to the double coding sessions, they first had to be educated on

the subject. Then a test run was done, by splitting the data set into one for training, containing 9 of the participants, and one for the coding containing the other 25. In order to test the validity of how coherent the researcher and the outside party were, Spearman's Rho, as explained by *Spearman's Rank-Order Correlation* [75], was calculated for each question. Each question had a Rho of above 0.95, meaning that the two coders had a strong association of the direction of the questions.

Exploratory Measures
For the exploratory measures we wanted to see which of the two conditions was more enjoyable for the participants. This was explored in the demographics as well. The questions for this were also drawn up from the IMI, as per their sub-scale for measuring the participants' interest/enjoyment.

4.1.4. Participants

The participants had to complete two interactions with a virtual agent, which was available on a local environment on the laptop of the researcher. While some were done in person, most had to be done through remote access control through TeamViewer[5] as circumstances were preventing the researcher from traveling. Of all the participants, 18 performed over TeamViewer, while the other 16 did it in person. For both conditions, the researcher left the room after the explanation and could be contacted on their phone if the participant had any questions. In total, 34 participants completed the experiment. The rationale behind the sample size is found using the paper written by Brinkman [12], who says that for a two-tailed *t*-test, the sufficient participant size is 34 and to obtain .80 power when using a medium effect size at the standard .05 alpha error probability. The only prerequisites were that participants could speak both Dutch and English to engage with both the system and the questionnaires properly. Participants were recruited from the public. No participants were familiar with the system beforehand.

In Table 4.1, an overview of the participants can be seen. In total, 26 males, 7 females, and 1 non-binary person were interviewed. Two other participants did not complete the experiment due to technical errors. Their input was not taken into account. For the knowledge questions: some of the participants opted to leave some of the open-ended questions empty, with them citing that they could not think of an answer. These participants were rewarded 0 points for the answer, and the questionnaires were still used for the analysis.

Participants			
Statistics	Male	Female	Non-binary
number	26	7	1
average age	25	28	25
standard deviation age	1.67	4.91	-
range age	20-29	23-38	-

Table 4.1: Statistics of the participants

4.1.5. Procedure

The experiment's procedure could go one of two ways for any participant. The two different paths are outlined in Figure 4.3 below. First, the participant read and filled out the consent form. They then received an explanation of how the procedure would work in detail. This included information on the two sessions they had with the chatbot, as well as a video explaining the different phases of the 5-phase model to them. Examples used during the video differed from the scenario used during experimentation. They could ask questions if anything was unclear and could ask to watch the video again as well.

Then, the participant randomly got assigned one of two conditions, either use the explanation sheet or receive immediate feedback from the system. In the case of the former, participants were only allowed to open the explanation sheet after the program had started, thereby preventing any unsanctioned usage that could influence their results.

When the participants finished their interaction with the chatbot they were asked to notify the researcher. They were then given the first questionnaire measuring both the perceived usefulness of the advice material and their knowledge of the 5-phase model. After filling out a questionnaire, these

[5] https://www.teamviewer.com/nl/

participants were then asked to go through another conversation this time having the direct feedback to guide them. The other group used these conditions in reverse order. By letting both groups do both conditions, it is believed that the biases could be counterbalanced against each other. To reduce bias even further, a double-blind approach was used to randomize which participant started with what condition. To ensure that neither participant nor researcher was biased by prior knowledge or expectations. These steps are shown in Figure 4.2 below.

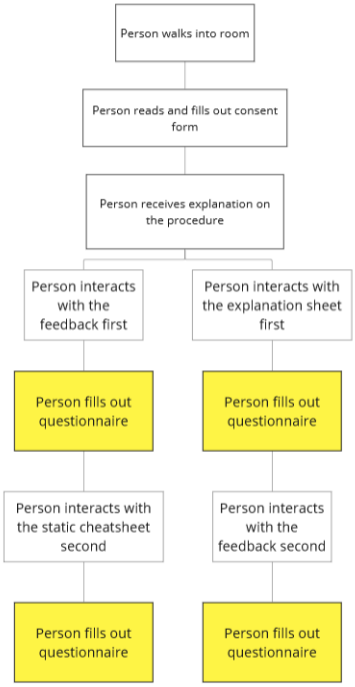


Figure 4.2: The basic setup of the experiment

4.1.6. Data preparation and statistical analysis

The data concerning the two hypotheses measuring perceived usefulness and knowledge on the 5-phase model were each analysed using a *t*-test. Both used a paired samples students' *t*-test with a standard error of 5 percent.

For the data preparation, knowledge of the 5-phase model was measured through questions and thus would be scored on a rubric, the scores of which would then undergo the *t*-test. The perceived usefulness questions have a built-in calculation that converts the participants' answers from the Likert-scale to a numeric value. For reliability, we used Cronbach's alpha as found in Peterson [65].

To test the reliability of the knowledge test gradings, the data was processed by the researcher and an outside observer, with a background in mechanical engineering. For knowledge, the 34 participants were split up in a training sample of 9 and a test sample of 25. However, the test sample was taken at random, which caused the final test samples to be unbalanced. In total, there were 15 people graded who started with immediate feedback, and 10 who started with the explanation sheet.

To evaluate the IMI results, we looked at several papers using IMI. The results of which are in Table 4.2. Shown is the interest/enjoyment sub-scale as that was the common denominator between all papers. The only paper to use perceived usefulness as a sub-scale was Deci et al. [23].

All data was uploaded to 4TU.research data repository [6] and can be accessed from there.

Finally, due to human error, the exploratory measures were asked of only 25 of 34 participants.

[6] https://data.4tu.nl/9c68a82e-ad6c-420b-88dd-2e86ec729ffb

		Items	Interest / Enjoyment	n
		Immediate feedback system for children's helpline counselors	4.7	25
		Computer-based mathematics learning	4.25	22
		Promote internalization	2.80	48
		Augmented feedback on walking speed	6.3	18
		Remotely administered gamified Stop-Signal Task	5.15	30
		VR for motor skill	6.3	95

Table 4.2: Our IMI compared to other research

4.2. Results

In Table 4.3, all the results of the various questionnaires are shown. Each subsection will go into greater detail discussing said results.

Item	Mean		SD		N	p-value	Cohen's d
	Feedback	Sheet	Feedback	Sheet			
Perceived Usefulness	33.76	32.5	6.72	6.16	34	0.16	0.2
Knowledge	7.2	7.0	1.59	1.42	25	0.64	0.1
Enjoyment	32.7	30.8	8.64	9.05	25	0.26	0.2

Table 4.3: The results of the various questionnaires

4.2.1. Perceived Usefulness

In Figure 4.3 there is a box-plot showing how the data was distributed, the data being the perceived usefulness results of the questionnaire, showing that the means for perceived usefulness were quite close with the immediate feedback taking a slight edge, as well as a higher variance.

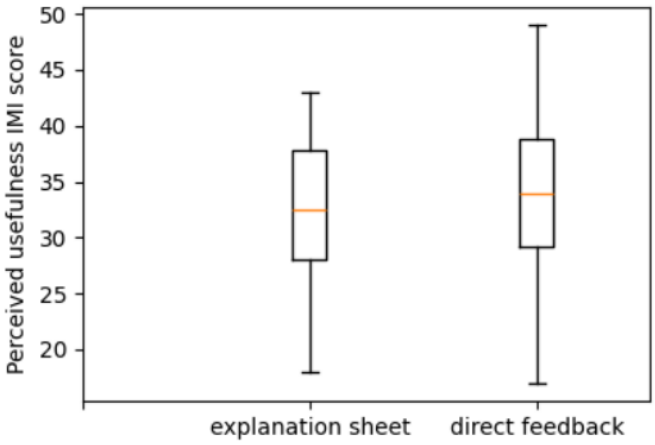


Figure 4.3: Box-plot of the perceived usefulness

According to Table 4.3 the results of perceived usefulness seem roughly similar between the immediate feedback and explanation sheet conditions. The means and standard deviation differed only slightly, with immediate feedback taking a slight, advantage (M = 33.76, SD = 6.72) compared to the explanation sheet (M = 32.5, SD = 6.16). From the paired t-test, the results were inconclusive, $t(33)$ = 0.81, p = 0.16. This means we cannot say for certain that any of the two methods is perceived as significantly more useful than the other, but we can neither say that the opposite is true. The question remains open for now. For the effect size of the difference between the two groups, Cohen's d was calculated. For perceived usefulness, the result was 0.2. For reliability, we calculated Cronbach's alpha, which for perceived usefulness was 0.66. This result is slightly below the average according to

Peterson [65].

4.2.2. Knowledge

According to the results in Table 4.3, similar to the perceived usefulness, there is no obvious gap in terms of results for both the immediate feedback (M = 7.2, SD = 1.59) and the explanation sheet (M = 7.0, SD = 1.42). The results of the t-test were inconclusive, $t(24) = 0.34$, $p = 0.64$. This means that the question remains open, neither of the two conditions was significantly better than the other, however, we also cannot say that they are exact equals in terms of improving knowledge. For the effect size of the difference between the two groups, Cohen's d was calculated. For knowledge, the result was 0.1.

4.2.3. Double Coding

Double coding was used for reliability and Spearman's Rho was calculated to check how coherent both parties were with one another, with each question having a Spearman's Rho of above 0.995.

4.2.4. Exploratory Measures

Due to the aforementioned human error, 17 males (68%), 7 females (28%), and 1 non-binary person (4%) joined for the exploratory measures, giving us a total of 25 people.

Figure 4.4 is a box-plot showing how the enjoyment results of the questionnaire were distributed. With direct feedback having a higher mean than the explanation sheet. Once again, the effect size was calculated using Cohen's d. For the enjoyment, the result was 0.2. The statistics can be seen in Table 4.3. For reliability, the Cronbach's alpha for enjoyment was lower than that of perceived usefulness, with an alpha of 0.59. When looking at the IMI scores from other papers in Table 4.2, taking the average of the other papers, we find that their score per point averages out to 4.96, which means that our 4.70 is slightly below it.

There is little difference between the two data sets for immediate feedback (M = 32.7, SD = 8.64) and the explanation sheet (M = 30.8, SD = 9.05). The t-test gave us once again inconclusive results, $t(24) = 0.80$, $p = 0.26$. This is a similar result to what we had for the primary measures and the question remains open as no proper conclusions can be drawn.

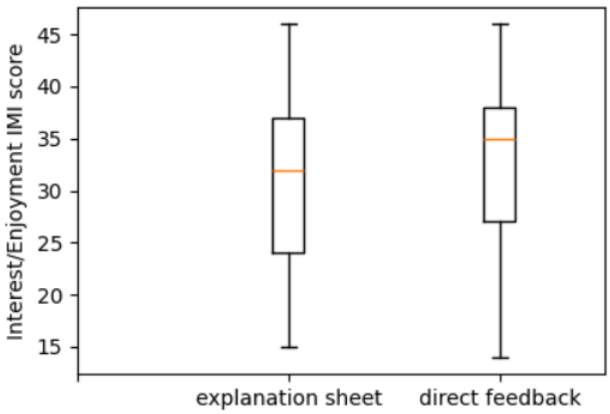


Figure 4.4: Box-plot of the enjoyment

With all the inconclusive data, we would like to explore the dataset more. While we cannot show that immediate feedback is better or worse than an explanation sheet, we can look into immediate feedback on its own. By taking the data and doing a between-subject study, taking only half of the data from the first condition where the participants interacted, we hope to better understand our data set. Looking at the box-plot in Figure 4.5, we can see that overall direct feedback ranks a little higher when only observing the dataset from this angle. Furthermore, in Figure 4.6, it can be seen that the feedback

system participants have a few high scorers in both perceived usefulness and knowledge.

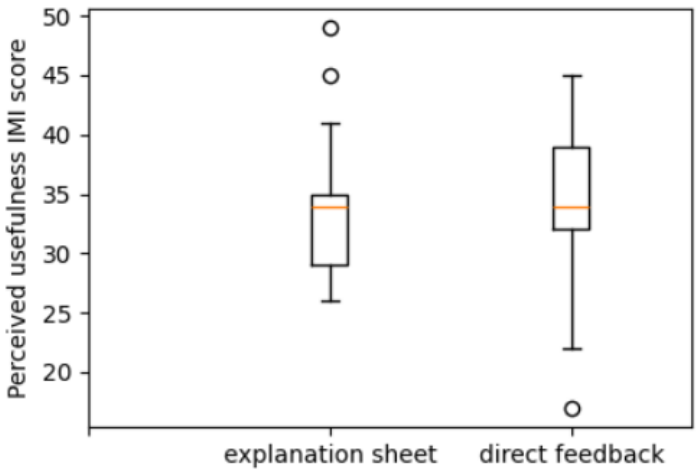


Figure 4.5: Box-plot of the perceived usefulness, looking only at the first time a participant interacted with a given condition

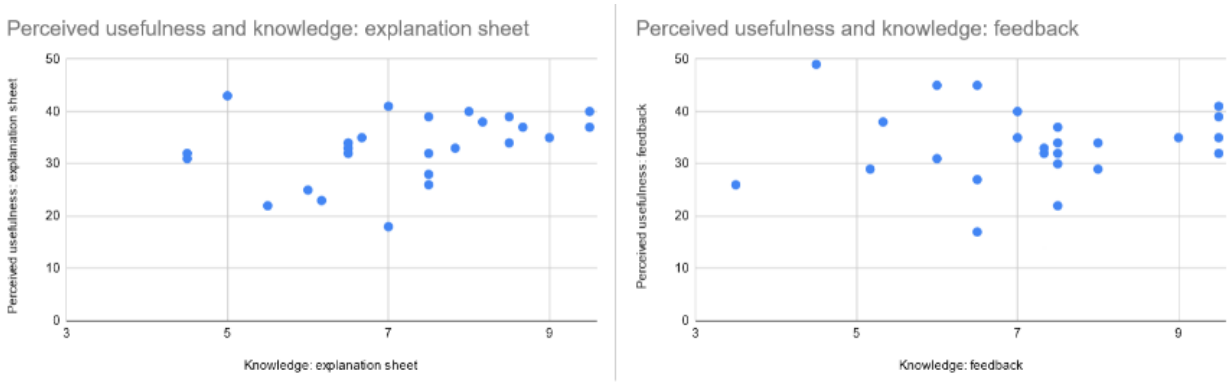


Figure 4.6: Two scatter plots of the perceived usefulness and knowledge, for each condition

On the other hand, when looking at box-plot in Figure 4.7, the explanation sheet seems to perform better than direct feedback on knowledge. When only looking at the first time a participant interacted with a given condition, the mean is visibly higher for the explanation sheet. Direct feedback also had a greater variance in the score.

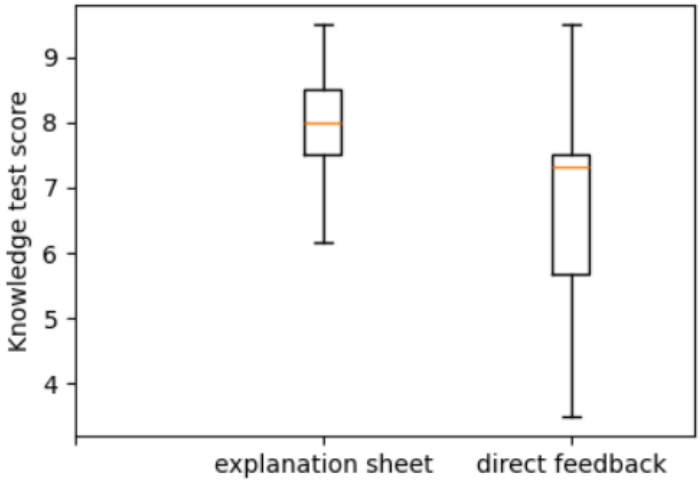


Figure 4.7: Box-plot of the knowledge score, looking only at the first time a participant interacted with a given condition

4.3. Discussion

For all the hypotheses, we expected there to be a statistically significant difference between the two conditions. For all of our *t*-tests, however, the results were inconclusive, meaning that we could not find whether there was a difference between receiving immediate feedback or using the explanation sheet next to them.

We originally used Brinkman [12] described G*power tool to calculate that we needed 34 participants for a medium effect size of 0.8 for our experiment. Subsequently, upon encountering a Cohen's d of 0.2, we revisited the calculation using the tool once again. This reevaluation revealed that the necessary participant count would need to be 200 to reach an effect size of 0.8.

When discussing perceived usefulness, challenges arise. No conclusion is evident from the data alone. For a solution, a possible avenue could be the work done by Phillips [66], who notes that the longer a session with a chatbot is, the more engaging it is. What constitutes as a long enough session for engagement differs from context to context, with Hew et al. [40] finding that, on average, after 4 minutes a participant is properly engaged with the chatbot. We wonder if the same could be said for the perceived usefulness of the feedback. In our study, any given conversation usually lasted between 12 to 15 minutes, with the shortest one being 6 minutes and the longest conversation taking 18. It could be that by altering the feedback towards a shorter conversation, that perceived usefulness could have increased.

When looking at the numeric values of the IMI and at Figure 4.4, participants using feedback on average gave the system a score of 33.8, which on a 7-question IMI sub-scale is a score of 4.82 per point. For the explanation sheet, the average score per question is 4.64. Both are higher than anything found in other experiments measured with IMI, such as Deci et al. [23] and Choi, Mogami, and Medalia [17]. The same is also true for the IMI scores for interest/enjoyment, which were 4.67 for immediate feedback and 4.39 for the explanation sheet.

As for knowledge, some takeaways were interesting when looking at the exploratory analysis. The explanation sheet's score is higher than those of immediate feedback. A possibility for this is that people actively seeking information causes them to better absorb it, perhaps similar to the Ask-Tell-Ask communication strategy as devised by French et al. [30] which is a learner-centered approach, where the learner can reflect on their own capabilities before receiving feedback.

Taking everything into consideration, the sub-research question can not be suitably answered. Participants would note to the researcher that they found the system promising and interesting, but not particularly useful in its current implementation. They mostly cited its potential and cited that the feedback part was working properly, but that it seemed aimed at trying to teach them how to operate the chatbot, rather than to learn how to hold a proper conversation.

4.4. Limitations

This study was set to find out what feedback system is useful for possible children helpline counselors. From conversations with De Kindertelefoon as well as from observations made during research, it is hard to pinpoint what an average Kindertelefoon counselor is. However, of the participants, 33 were between age 20 and 29. This, in concert with an overwhelming amount of the participants being male, makes the group of participants not particularly representative of the more varied Kindertelefoon counselor.

Also, as was mentioned in the exploratory measures, when converting the experiment to a between-subject study, it seemed that the explanation sheet performed a bit better. One possible reason is that the act of using the explanation sheet causes the user to be more active than waiting for the immediate feedback. Perhaps text prompts with contextual information could be used to simulate a similar experience.

One limitation can be found with the chatbot itself. After filling in the questionnaires, many participants mentioned that they had issues with getting the chatbot to understand them. Saying that the wording required to progress through the conversation was too specific. Others described that they saw the chatbot as a puzzle and ignored the advice material, trying to ”game” the chatbot into giving them a response they considered appropriate. Another limitation of the chatbot was a light tendency to crash. 5 of the 34 participants had messaged the researcher during a conversation at least once, alerting them of an actual crash of the system. These crashes were seemingly random with the condition used not affecting it.

Finally, another limitation seemed to be a sensory or memory overload of the participants. At the start of the examination, a large amount of information is given to the participants. They receive information on the study, their participation, the basic workings of the bot, and of course the 5-phase model. Participants occasionally asked for a repetition of information, citing that they were overwhelmed. Perhaps this overloading of information made it harder for them to concentrate on the conversation with Lilobot, thus impacting their experience.

5

Conclusion

In the final chapter, the answers to the different research questions are summarized and the limitations, as well as the contributions of the system, will be discussed.

5.1. Conclusion

This research aimed to answer the following research question:

What feedback-system design is useful and feasible for training children's helpline volunteers?

We split these up into the following sub-questions:

What are the design factors and concerns for the feedback system?
Feedback can be both immediate and delayed. Immediate feedback can be in the beginning when the trainees still have much to learn and when the amount of errors made is at its highest. This immediate feedback could be given with a neutral disposition. Meaning that it should not be given in an accusatory manner, but rather as a matter-of-fact. This is because De Kindertelefoon wants to keep their would-be counselors motivated and not overwhelm them with negative feedback. De Kindertelefoon wants to promote positive feedback often believing it to be the most important for information retention, as some of the found literature reflects this as well, such as Metcalfe, Kornell, and Finn [61] and Van der Kleij, Feskens, and Eggen [79]. As the trainee moves along further in the training program, their skills and knowledge mature and the feedback should reflect this.

How could feedback be integrated in a conversational agent training environment to attain a high perceived usefulness of the system as well as the knowledge on the 5-phase model?
Feedback has been integrated through direct feedback, both positive and corrective. Feedback is based on a participant's current input and the input that came prior. Delivered to counselors-in-training as part of the UI, the system's feedback guides the participants through a handcrafted path, which is based on the Beliefs and Desires of the chatbot. The feedback helps the participants navigate through the conversation, while also teaching the participants the intricacies of how, when, and why to use certain parts of the 5-phase model.

How useful do trainees find the feedback, and how much knowledge do they gain?
People stated they saw potential for usefulness to the system, but that its current form was lacking. From the inconclusive data, we cannot say whether people prefer the feedback system over having an explanation sheet with information on all 5 phases next to them, meaning that the question remains open for now. A part of it could be that a person's user agency was more important than the direct feedback received from the system. Some people are more attuned to finding the answers themselves within the explanation sheet condition than trying to figure things out from the direct feedback condition. Furthermore, there seemingly was no difference in the gaining of knowledge on the 5-phase model between the two conditions, making the feedback system as apt as the explanation sheet, and possibly a good addition to it.

5.2. Limitations

We recognize that there are several limitations to the work provided.

The planned experiment was a within-subject study, which took a significant portion of time to set up and execute, but allowed us to use fewer participants than if we performed a between-subject study. Making it a within-study had the added benefit of giving more samples with which we could perform double-coding for the grading, allowing us to split up all of our test subjects into a training sample and a test sample. This split was done randomly in an attempt to reduce bias, but the random sampling did cause the feedback system condition to be represented more than the explanation sheet condition (15 versus 10 respectively). This is in addition to the recalculated needed number of participants, which turned out to be 200 instead of 34. Also, the age range of participants (20-29 for males, and 23-38 for females) might have been too limited to properly represent the population.

Unfortunately, many auxiliary statistics were not recorded which, in hindsight, could have proven to be useful. Hew et al. [40] for example, recorded how long their participants interacted with the chatbot, which, for us, could have given us insights into what method might seem more intuitive for a participant. Another statistic missed was checking whether or not the participant completed the scenario, and seeing which condition netted the highest results. In addition, we could have kept track of which phase the participant ended up in, to see which condition performed best. Many participants did note that the second time through they had an easier time navigating the bot, but hopefully the counterbalance would have weeded out this bias. In addition, during the data analysis, we were missing more open-ended questions where participants could have shared some of their insights into the system. While it would have cost more time, it would have allowed for stronger exploration of the data. But we also recognize that it would have been rather costly in terms of time and that not doing it did allow for more exploratory analyses.

Finally, due to extraneous circumstances, the collection of data was altered during the process. Forcing a switch from real-life participation on the researcher's laptop to allowing the participants to interact with the chatbot through remote-access-control. The researcher being present for an explanation of the system and the procedure could have impacted how well people absorbed the information given. However, remote-access-control did make the recruitment process easier, so when a participant suddenly dropped out, an alternative participant could be found without much difficulty.

5.3. Contributions

The research shows the possible strength of feedback in a practical manner. The practicality was key here, as the goal was to make a useful system outside of academic pursuits as well. From the results, it is apparent that the feedback does not resonate with people more than the classical approach of an explanation sheet. However, the results are not deterring either as they were close to said classical approach. Furthermore, from the focus group and experiment, it was learned that people do find the idea interesting. As the results were non-conclusive, future work could improve what was done.

From an academic point of view, the work mostly provides a launching pad to do more in-depth analyses in the future, as well as see what the effects are of combining the two conditions. The biggest contribution is the actual feedback system itself. As far as our knowledge is the first time such a system has been implemented for a chatbot in a hotline training simulation.

We have also set up a literature review covering our design guidelines for feedback and a feedback system, which covers how we believe it can promote agency, knowledge, and perceived usefulness. Other related topics include self-efficacy and the Self-Determination Theory.

Other contributions are our questionnaires and the insights on the effectiveness of immediate feedback. We hope that others working in a similar setting or field can use our results from both the IMI and the experiment as a whole.

5.4. Future Work

First and foremost, the chatbot should be improved. A linchpin in both development and testing, the current simplicity of the machine and its rather stringent curation of words makes the experience more about teaching a participant how to navigate the bot, rather than teaching them about the 5-phase

model. The bot must be expanded to take on more varied inputs, and to add to that, the bot should also be expanded to have more beliefs and desires. Though chatbots will continue to be used in a variety of fields such as music as shown by Garayzar-Cristerna and Luna-Ramirez [33], finances as shown by Sugumar and Chandra [76], and mental health as shown by Galindo et al. [32], this technology will increase over time allowing for more powerful tools.

Other similar work includes expanding the bot to breach more subjects than merely bullying. Adding more subjects beyond bullying can also provide more information and more practical training for De Kindertelefoon.

For possible improvements made to the feedback system itself, the system could be extended to include more detailed feedback specific to the user. In the current system, only the user's current input and previous state are used to determine the feedback, while better feedback could be given based on the context of the entire conversation. For example, questions about loved ones that the counselor-in-training asked earlier in the conversation can be used later in the conversation, the feedback system could make use of this by reminding the user what it had already said prior. An additional improvement for the feedback system could also be to give more detailed feedback upon the counselor-in-training's request. If a person is in the middle of a training session, they could request feedback on their current progress, including the steps they have taken and what they have missed.

Furthermore, the system could be extended to include updated and in-depth delayed feedback. Said feedback could go into greater detail on what the user said and how well that input works in any given context. Furthermore, delayed feedback could also be used to expand the knowledge on the disparity between immediate and delayed feedback, perhaps similar to Masantiah, Pasiphol, and Tangdhanakanond [58]. From the repeated conversations, we did hear from participants that, going through it a second time, they found it easier to navigate the chatbot.

The system could also be ”upgraded” beyond a mere chatbot and start using AI voice modulation to recreate a phone call with a child, which is the other commonly used method of communication with De Kindertelefoon. With this emergent technology, feedback could be provided on prosodic methods such as tone and pitch, and it could provide a more realistic training example for the volunteers, akin to the work done by Wang et al. [81], which automatically recommended modulation examples for public speeches, with immediate feedback as well to guide further improvement.

Finally, a way to promote even more agency would be to allow the users to decide when they ask for feedback, which could be particularly interesting for researching how different demographics handle feedback, as was done similarly by Reiser, Van Vreede, and Petty [68]. This could be done in conjunction with De Kindertelefoon as well, as they have a diverse number of counselors.

5.5. Final Remarks

In this thesis, we have created an instant feedback system to help train counselors to work at a children's helpline. The evaluation shows that the added feedback system on top of an already existing chatbot system did not give conclusive results on neither the perceived usefulness of the user, nor on the knowledge gained on the 5-phase conversational model, used by the children's helpline. In the future, more research can be done to make the feedback system suitable for a supportive tool, which would work in conjunction with more classical approaches. Finally, people's interest in the subject could indicate that the work might be more in line with people's style of learning and a more refined version could prove to be more fruitful.

Bibliography

[1] Jette Ammentorp et al. “The effect of training in communication skills on medical doctors’ and nurses’ self-efficacy: A randomized controlled trial”. In: *Patient education and counseling* 66.3 (2007), pp. 270–277.

[2] Craig G Anderson et al. “Failing up: How failure in a game environment promotes learning through discourse”. In: *Thinking Skills and Creativity* 30 (2018), pp. 135–144.

[3] Albert Bandura. “Self-efficacy mechanism in human agency.” In: *American psychologist* 37.2 (1982), p. 122.

[4] Albert Bandura. “The explanatory and predictive scope of self-efficacy theory”. In: *Journal of social and clinical psychology* 4.3 (1986), pp. 359–373.

[5] Stuart Beattie et al. “The role of performance feedback on the self-efficacy–performance relationship.” In: *Sport, Exercise, and Performance Psychology* 5.1 (2016), p. 1.

[6] Abir K Bekhet and Jaclene A Zauszniewski. “Methodological triangulation: An approach to understanding data”. In: *Nurse researcher* 20.2 (2012).

[7] Anique de Beyn. *In gesprek met kinderen: de methodiek van de kindertelefoon*. NIZW, 2003.

[8] Gert Biesta and Michael Tedder. “Agency and learning in the lifecourse: Towards an ecological perspective”. In: *Studies in the Education of Adults* 39.2 (2007), pp. 132–149.

[9] DR Billings. “Efficacy of adaptive feedback strategies in simulation-based training”. In: *Military Psychology* 24.2 (2012), pp. 114–133.

[10] C Brame. “Writing good multiple choice test questions”. In: *Center for Teaching Vanderbilt University* (2013).

[11] Petter Bae Brandtzaeg and Asbjørn Følstad. “Why people use chatbots”. In: *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*. Springer. 2017, pp. 377–392.

[12] Willem-Paul Brinkman. “Design of a questionnaire instrument”. In: *Handbook of mobile technology research methods*. Nova Publishers, 2009, pp. 31–57.

[13] Simon Brownhill. “Asking key questions of self-reflection”. In: *Reflective Practice* 23.1 (2022), pp. 57–67.

[14] Christian Burgers et al. “How feedback boosts motivation and play in a brain-training game”. In: *Computers in Human Behavior* 48 (2015), pp. 94–103.

[15] John M Carroll. *Making use: scenario-based design of human-computer interactions*. MIT press, 2003.

[16] Dympna Casey and Kathy Murphy. “Issues in using methodological triangulation in research”. In: *Nurse researcher* 16.4 (2009).

[17] Jimmy Choi, Tamiko Mogami, and Alice Medalia. “Intrinsic motivation inventory: an adapted measure for schizophrenia research”. In: *Schizophrenia bulletin* 36.5 (2010), pp. 966–976.

[18] Roy B Clariana, Daren Wagner, and Lucia C Roher Murphy. “Applying a connectionist description of feedback timing”. In: *Educational Technology Research and Development* 48.3 (2000), pp. 5–22.

[19] Fabio Clarizia et al. “E-learning and industry 4.0: A chatbot for training employees”. In: *Proceedings of Fifth International Congress on Information and Communication Technology: ICICT 2020, London, Volume 2*. Springer. 2021, pp. 445–453.

[20] Iris Cohen, Willem-Paul Brinkman, and Mark A. Neerincx. “Effects of different real-time feedback types on human performance in high-demanding work conditions”. In: *International Journal of Human-Computer Studies* 91 (2016), pp. 1–12. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2016.03.007. URL: https://www.sciencedirect.com/science/article/pii/S1071581916000392.

[21] Matthieu Courgeon et al. “MACH: My automated conversation coach”. In: Dec. 2014. DOI: 10.1145/2493432.2493502.

[22] Edward L Deci et al. “Conceptualizations of intrinsic motivation and self-determination”. In: *Intrinsic motivation and self-determination in human behavior* (1985), pp. 11–40.

[23] Edward L Deci et al. “Facilitating internalization: The self-determination theory perspective”. In: *Journal of personality* 62.1 (1994), pp. 119–142.

[24] Damiaan Denys. *Het tekort van het teveel: de paradox van de mentale zorg*. Singel Uitgeverijen, 2020.

[25] Michael Emmison and Susan Danby. “Troubles Announcements and Reasons for Calling: Initial Actions in Opening Sequences in Calls to a National Children’s Helpline”. In: *Research on Language and Social Interaction* 40.1 (2007), pp. 63–87. DOI: 10.1080/08351810701331273. eprint: https://doi.org/10.1080/08351810701331273. URL: https://doi.org/10.1080/08351810701331273.

[26] Guido A. Entenberg et al. “Using an Artificial Intelligence Based Chatbot to Provide Parent Training: Results from a Feasibility Study”. In: *Social Sciences* 10.11 (2021). ISSN: 2076-0760. DOI: 10.3390/socsci10110426. URL: https://www.mdpi.com/2076-0760/10/11/426.

[27] Michael L. Epstein, Beth B. Epstein, and Gary M. Brosvic. “Immediate Feedback during Academic Testing”. In: *Psychological Reports* 88.3 (2001). PMID: 11508040, pp. 889–894. DOI: 10.2466/pr0.2001.88.3.889. eprint: https://doi.org/10.2466/pr0.2001.88.3.889. URL: https://doi.org/10.2466/pr0.2001.88.3.889.

[28] Lauren Eskreis-Winkler and Ayelet Fishbach. “Not learning from failure—The greatest failure of all”. In: *Psychological science* 30.12 (2019), pp. 1733–1744.

[29] Mary Forehand et al. “Bloom’s taxonomy: Original and revised”. In: *Emerging perspectives on learning, teaching, and technology* 8 (2005), pp. 41–44.

[30] Judith C French et al. “Targeted feedback in the milestones era: utilization of the ask-tell-ask feedback model to promote reflection and self-assessment”. In: *Journal of Surgical Education* 72.6 (2015), e274–e279.

[31] Ruben Fukkink and Jo Hermanns. “Counseling children at a helpline: chatting or calling?” In: *Journal of community psychology* 37.8 (2009), pp. 939–948.

[32] Mauricio J Osorio Galindo et al. “E-Friend: A Logical-Based AI Agent System Chat-Bot for Emotional Well-Being and Mental Health”. In: *Deceptive AI: First International Workshop, DeceptECAI 2020, Santiago de Compostela, Spain, August 30, 2020 and Second International Workshop, DeceptAI 2021, Montreal, Canada, August 19, 2021, Proceedings 1*. Springer. 2021, pp. 87–104.

[33] Arantxa Garayzar-Cristerna and Wulfrano Arturo Luna-Ramirez. “ADAGIO, a BDI Music Recommender Telegram Chatbot”. In: *Science and Information Conference*. Springer. 2023, pp. 175–184.

[34] Talip Gonulal and Shawn Loewen. “Scaffolding technique”. In: *The TESOL encyclopedia of English language teaching* (2018), pp. 1–5.

[35] Sharon Grundmann. *A BDI-based Virtual Agent for Training Child Helpline Counsellors*. 2022. DOI: https://doi.org/10.4121/17371919. URL: http://resolver.tudelft.nl/uuid:f04f8f0b-9ab9-4f1c-a19c-43b164d45cce.

[36] Bahar Gün. “Quality self-reflection through reflection training”. In: *ELT journal* 65.2 (2011), pp. 126–135.

[37] Thomas M Haladyna. “Developing Test Items for Course Examinations. IDEA Paper# 70.” In: *IDEA Center, Inc.* (2018).

[38] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. “Design and Evaluation of Explainable BDI Agents”. In: *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. Vol. 2. 2010, pp. 125–132. DOI: 10.1109/WI-IAT.2010.115.

[39] Matthew Jensen Hays, Nate Kornell, and Robert A Bjork. “The costs and benefits of providing feedback during learning”. In: *Psychonomic bulletin & review* 17.6 (2010), pp. 797–801.

[40] Khe Foon Hew et al. “Using chatbots in flipped learning online sessions: perceived usefulness and ease of use”. In: *Blended Learning: Re-thinking and Re-defining the Learning Process. 14th International Conference, ICBL 2021, Nagoya, Japan, August 10–13, 2021, Proceedings 14*. Springer. 2021, pp. 164–175.

[41] Edmund S. Higgins. “Is Mental Health Declining?” In: *Scientific American Mind* 28.1 (2017), pp. 20–22. ISSN: 15552284, 2331379X. URL: https://www.jstor.org/stable/24945571 (visited on 11/11/2022).

[42] Sanna Hilden and Kati Tikkamäki. “Reflective practice as a fuel for organizational learning”. In: *Administrative sciences* 3.3 (2013), pp. 76–95.

[43] Edward Howie et al. “Human–computer interface design can reduce misperceptions of feedback”. In: *System Dynamics Review: The Journal of the System Dynamics Society* 16.3 (2000), pp. 151–171.

[44] Magid Igbaria and Juhani Iivari. “The effects of self-efficacy on computer usage”. In: *Omega* 23.6 (1995), pp. 587–605.

[45] Seppo E Iso-Ahola and Charles O Dotson. “Psychological momentum: Why success breeds success”. In: *Review of general psychology* 18.1 (2014), pp. 19–33.

[46] *Jaarverslag 2019*. https://jaarverslag.kindertelefoon.nl/2019. Accessed: 2022-11-14.

[47] *Jaarverslag 2021*. https://jaarverslag.kindertelefoon.nl/2021. Accessed: 2022-11-14.

[48] Komal Joshi and Ram Lal Yadav. “A new hybrid approach for solving travelling salesman problem using ordered cross over 1 (ox1) and greedy approach”. In: *IJRET: International Journal of Research in Engineering and Technology* 4.05 (2015).

[49] Katherine A Karl, Anne M O’Leary-Kelly, and Joseph J Martocchio. “The impact of feedback and self-efficacy on performance in training”. In: *Journal of Organizational Behavior* 14.4 (1993), pp. 379–394.

[50] Toula Kourgiantakis, Karen M Sewell, and Marion Bogo. “The importance of feedback in preparing social work students for field education”. In: *Clinical Social Work Journal* 47.1 (2019), pp. 124–133.

[51] Kathleen M Krumhus and Richard W Malott. “The effects of modeling and immediate and delayed feedback in staff training”. In: *Journal of Organizational Behavior Management* 2.4 (1980), pp. 279–293.

[52] Gilbert Laporte. “The traveling salesman problem: An overview of exact and approximate algorithms”. In: *European Journal of Operational Research* 59.2 (1992), pp. 231–247.

[53] Pedro Larranaga et al. “Genetic algorithms for the travelling salesman problem: A review of representations and operators”. In: *Artificial intelligence review* 13 (1999), pp. 129–170.

[54] Lisa M Larson et al. “Development and validation of the counseling self-estimate inventory.” In: *Journal of counseling Psychology* 39.1 (1992), p. 105.

[55] Alexander Lidén and Karl Nilros. *Percieved benefits and limitations of chatbots in higher education*. 2020.

[56] Nikola Marangunić and Andrina Granić. “Technology acceptance model: a literature review from 1986 to 2013”. In: *Universal access in the information society* 14 (2015), pp. 81–95.

[57] David Markland and Lew Hardy. “On the factorial and construct validity of the Intrinsic Motivation Inventory: Conceptual and operational concerns”. In: *Research quarterly for exercise and sport* 68.1 (1997), pp. 20–32.

[58] Chutaphon Masantiah, Shotiga Pasiphol, and Kamonwan Tangdhanakanond. “Student and feedback: Which type of feedback is preferable?” In: *Kasetsart Journal of Social Sciences* 41.2 (2020), pp. 269–274.

[59] Jean McKendree. “Effective Feedback Content for Tutoring Complex Skills”. In: *Human–Computer Interaction* 5.4 (1990), pp. 381–413. DOI: 10.1207/s15327051hci0504_2. eprint: https://www.tandfonline.com/doi/pdf/10.1207/s15327051hci0504_2. URL: https://www.tandfonline.com/doi/abs/10.1207/s15327051hci0504_2.

[60] Conor Thomas McKevitt. “Engaging students with self-assessment and tutor feedback to improve performance and support assessment capacity”. In: *Journal of University Teaching & Learning Practice* 13.1 (2016), p. 2.

[61] Janet Metcalfe, Nate Kornell, and Bridgid Finn. “Delayed versus immediate feedback in children’s and adults’ vocabulary learning”. In: *Memory & cognition* 37.8 (2009), pp. 1077–1087.

[62] Berry O’Donovan et al. “What makes good feedback good?” In: *Studies in Higher Education* 46 (June 2019), pp. 1–12. DOI: 10.1080/03075079.2019.1630812.

[63] Bertram Opitz, Nicola K Ferdinand, and Axel Mecklinger. “Timing matters: the impact of immediate and delayed feedback on artificial language learning”. In: *Frontiers in human neuroscience* 5 (2011), p. 8.

[64] Harold Pashler et al. “When does feedback facilitate learning of words?” In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31.1 (2005), p. 3.

[65] Robert A Peterson. “A meta-analysis of Cronbach’s coefficient alpha”. In: *Journal of consumer research* 21.2 (1994), pp. 381–391.

[66] C Phillips. *Chatbot analytics 101: the essential metrics you need to track. Chatbots Magazine*. 2018.

[67] Alina Pommeranz et al. “Social acceptance of negotiation support systems: scenario-based exploration with focus groups and online survey”. In: *Cognition, Technology & Work* 14.4 (2012), pp. 299–317.

[68] Catherine Reiser, Victoria Van Vreede, and Elizabeth M Petty. “Genetic counselor workforce generational diversity: Millennials to Baby Boomers”. In: *Journal of Genetic Counseling* 28.4 (2019), pp. 730–737.

[69] Jennifer Robison, Scott McQuiggan, and James Lester. “Evaluating the consequences of affective feedback in intelligent tutoring systems”. In: *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE. 2009, pp. 1–6.

[70] Richard M Ryan and Edward L Deci. “Self-determination theory”. In: *Basic psychological needs in motivation, development, and wellness* (2017).

[71] Marisa Salanova, Isabel Martínez, and Susana Llorens. “Success breeds success, especially when self-efficacy is related with an internal attribution of causality”. In: *Estudios de Psicología* 33.2 (2012), pp. 151–165.

[72] R Keith Sawyer. “The role of failure in learning how to create in art and design”. In: *Thinking Skills and Creativity* 33 (2019), p. 100527.

[73] Mary Catherine Scheeler and David L Lee. “Using technology to deliver immediate corrective feedback to preservice teachers”. In: *Journal of behavioral education* 11.4 (2002), pp. 231–241.

[74] Trine Natasja Sindahl. *Chat Counselling for Children and Youth - A Handbook.* 2011.

[75] *Spearman’s Rank-Order Correlation*. https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php. Accessed: 2023-09-02.

[76] Moses Sugumar and Shalini Chandra. “Do I desire chatbots to be like humans? Exploring factors for adoption of chatbots for financial services”. In: *Journal of International Technology and Information Management* 30.3 (2021), pp. 38–77.

[77] Veronica A Thurmond. “The point of triangulation”. In: *Journal of nursing scholarship* 33.3 (2001), pp. 253–258.

[78] Robert J Vallerand. “Deci and Ryan’s self-determination theory: A view from the hierarchical model of intrinsic and extrinsic motivation”. In: *Psychological inquiry* 11.4 (2000), pp. 312–318.

[79] Fabienne M Van der Kleij, Remco CW Feskens, and Theo JHM Eggen. “Effects of feedback in a computer-based learning environment on students’ learning outcomes: A meta-analysis”. In: *Review of educational research* 85.4 (2015), pp. 475–511.

[80] Johan Von Wright. “Reflections on reflection”. In: *Learning and instruction* 2.1 (1992), pp. 59–68.

[81] Xingbo Wang et al. “Voicecoach: Interactive evidence-based training for voice modulation skills in public speaking”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–12.

[82] Weiner. *A growing psychiatrist shortage and an enormous demand for mental health services.* 2022.

[83] Bernard Weiner. “History of motivational research in education.” In: *Journal of educational Psychology* 82.4 (1990), p. 616.

Appendices

A. Focus Group

Focus group overview

The focus group was shown a story of a volunteer-in-training called Rob. To help him prepare, Rob was given a virtual training environment by De Kindertelefoon, where he can practice having a conversation with a virtual child who is bullied.

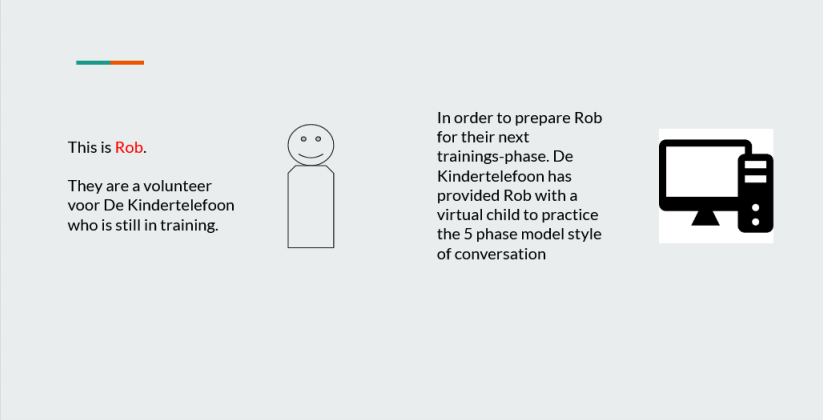


Figure 1: The focus group's introduction to Rob

Rob logs into the virtual environment and the first focus group scenario starts. Trainees develop skills over time and with that the amount and types of feedback and support they receive from their trainers changes as well. The focus group was shown two possibilities. One where the user, Rob, himself could tell the system his level of skill (going from beginner, to advanced, to veteran) and one where the system chooses it for him, based on his responses and the speed with which he gives them. The two options are visualised in Figure 2. The results are discussed at the end of chapter 2.2, as they tie into the topics of that section.

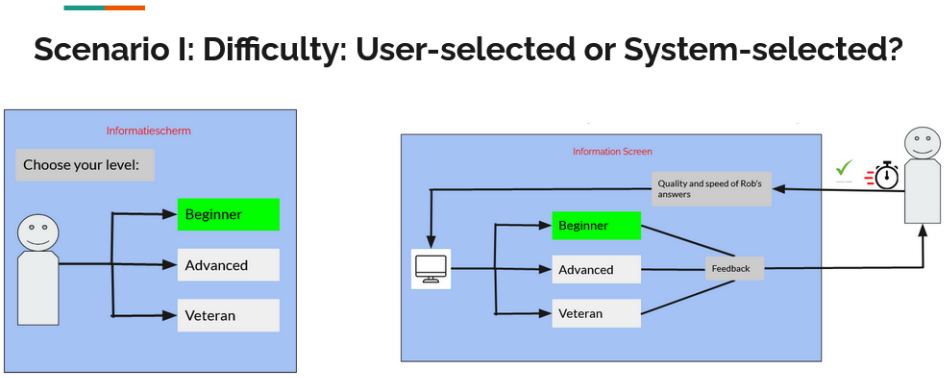


Figure 2: The two visual representations of different ways to determine difficulty. On the left the user "Rob" chooses the difficulty. On the right, the system takes his responses and the speed with which he gives them to select his difficulty for the next phase.

Rob received feedback either right after inputting an answer or after he had completed the simulated conversation. The scenario was called Immediate versus Delayed feedback. In Figure 3 you can see how it was presented. For the purpose of the focus group, the scenario was for us to learn when and how experienced trainers would give Rob feedback.

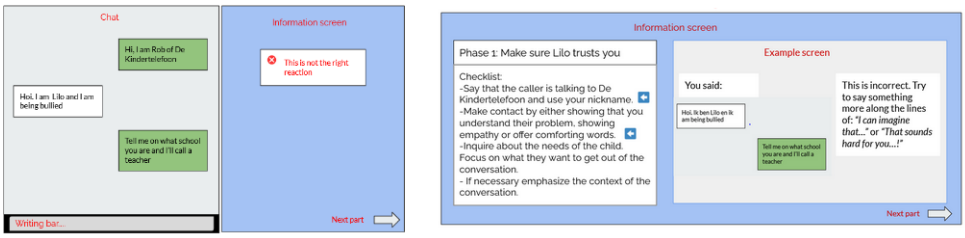
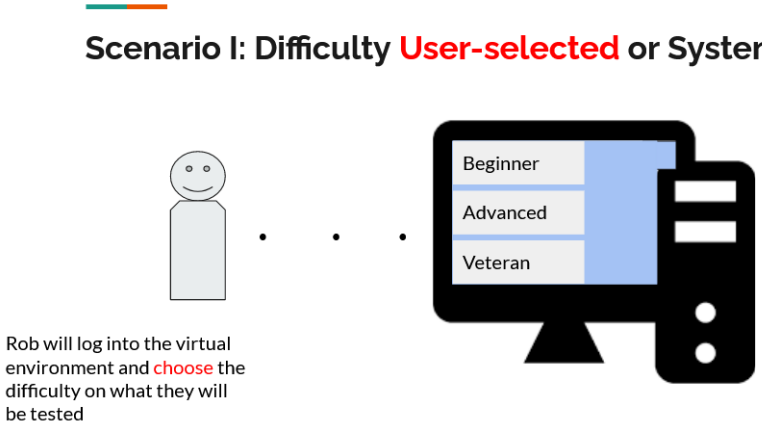
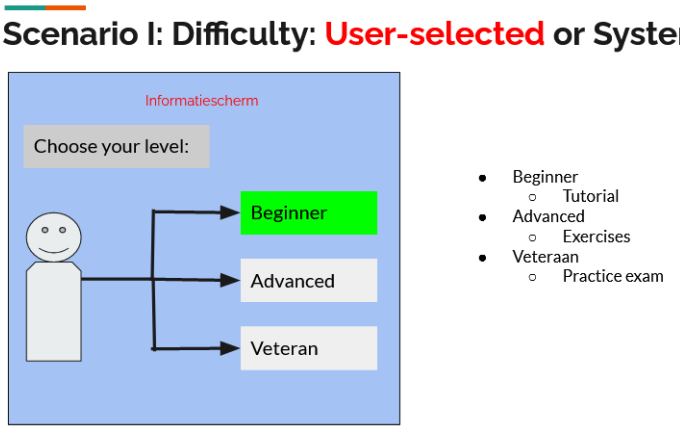


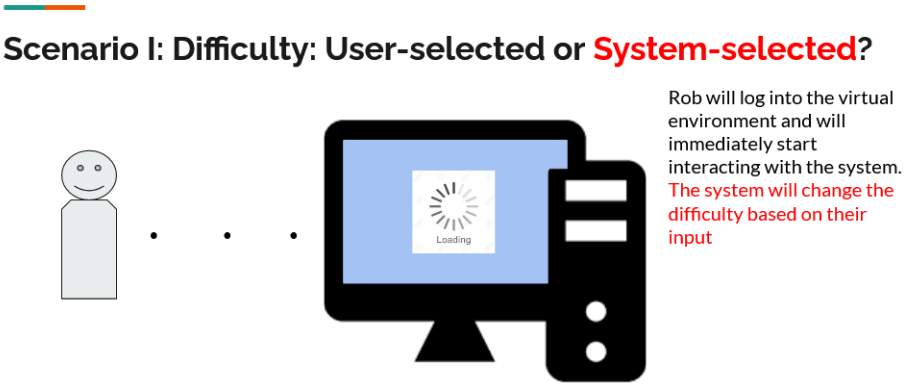
Figure 3: Immediate versus delayed feedback visualised. On the left, immediate feedback given to correct an erroneous input. On the right, an overview of the things ”Rob” did.



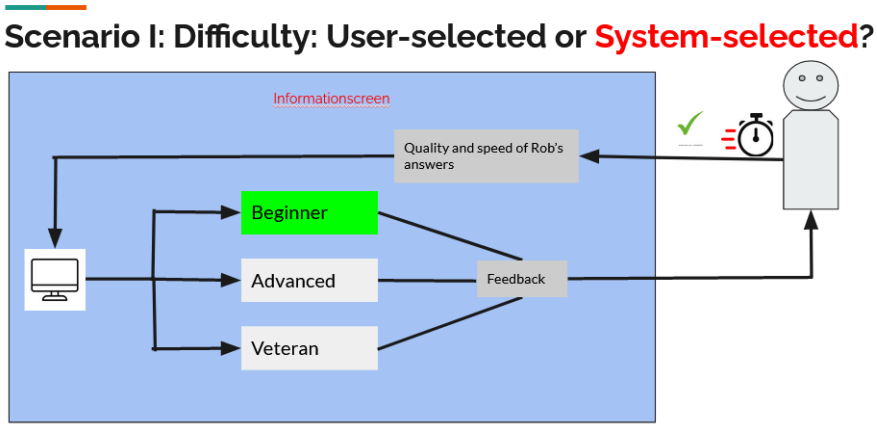
Focus Group Pictures
The first scenario gets explained. Rob will choose the difficulty of the system himself.



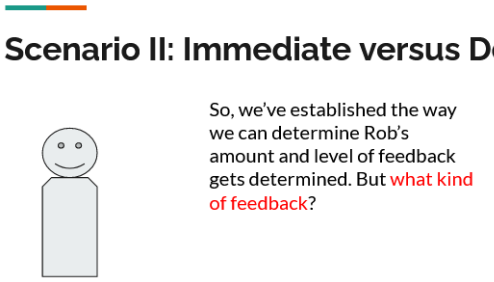
Rob chooses his difficulty and a short example of what one could expect at the different difficulties.



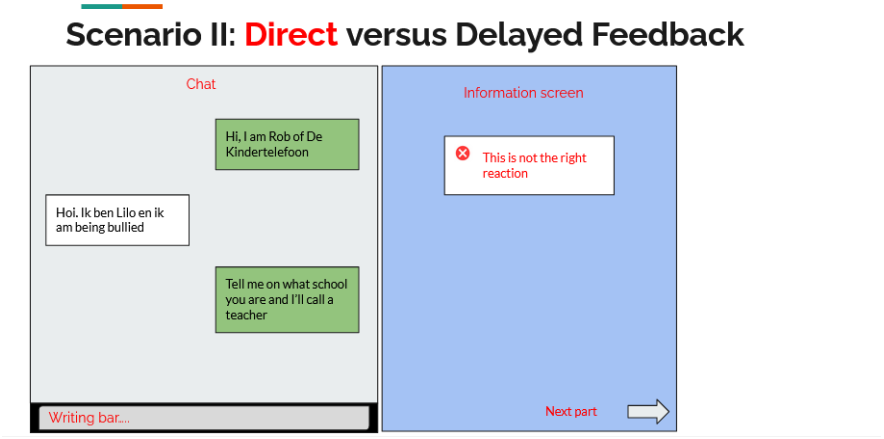
Introduction and explanation of the system-selected option.



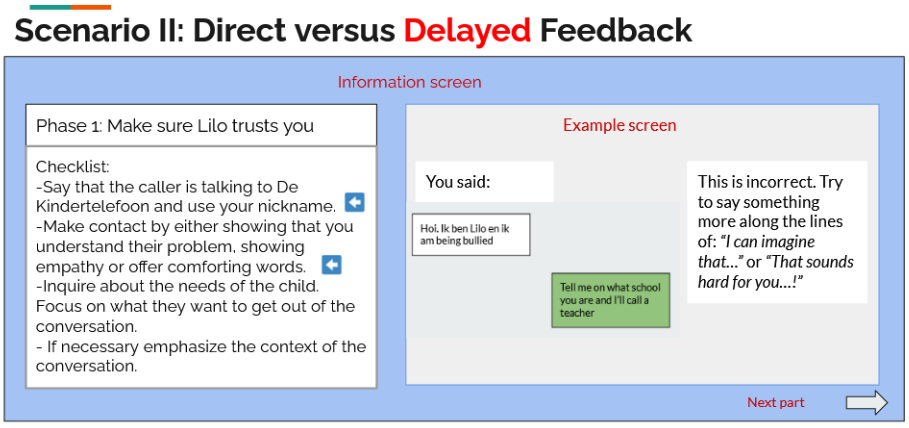
A visual representation of how the system-selected option works.



Introduction of the second scenario.



Visual representation of the direct feedback option.



Visual representation of the delayed feedback option.

Scenario II: Directe versus Delayed Feedback

The delayed feedback is more important than the direct feedback

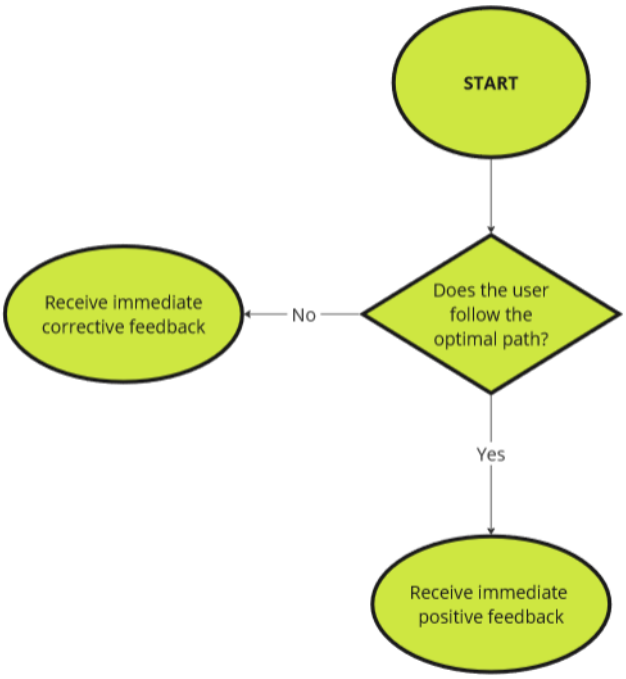
- Strongly Agree
- Agree
- Disagree
- Strongly Disagree

Slide showing the choice options for the focus group.

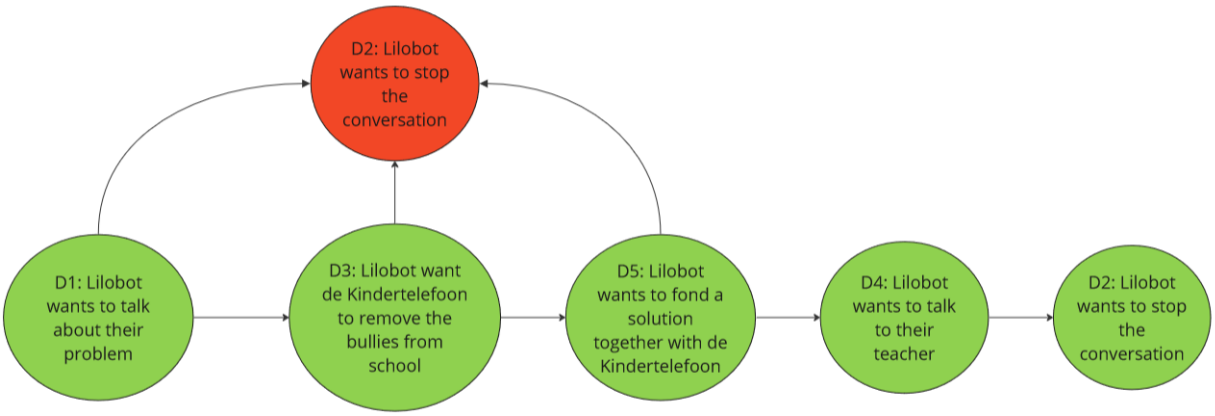
B. Design figures

Design Factor/Concern	Why	Design Idea
Immediate feedback both positive and corrective is important.	Immediate feedback is important to learn to overcome errors. Also the focus group mentioned that moments of success were key at keeping people motivated throughout training.	If a trainee makes a mistake, the system gives them the information needed to find the correct answer and then let them redo it. Corrective feedback can be given, but done so with tact.
If the system gives immediate feedback, which needs to be given with a neutral disposition.	We do not wish to disrupt the flow of the conversation or make the users worry too much about the feedback.	The feedback is not too friendly or aggressive, but delivered in a manner-of-fact disposition. Look to De Kindertelefoon for correct use of language.
Both types of feedback should not be given in an overwhelming amount.	The focus group voiced concern that too much feedback would disrupt the flow of the conversation and could overwhelm the trainee.	Immediate feedback can be given when a trainee gives incorrect input, but immediate positive feedback is only given at set intervals. Delayed feedback is succinct and can be given at the end of each phase.
The system could try to promote the user's agency, self-efficacy and autonomy.	From the gathered literature, it seems that agency helps both with motivation and retention.	Feedback could give contextual clues to the answer, instead of the answer outright.
Skills improve over time and that improvement could be tracked.	From the focus group, they would appreciate having a tool with which they can track a trainee's progress. Because, from their experience, people lack the ability to self-reflect and they lack the means to train people to gain that skill.	As the model is closely related to the 5-phase model, it can be worthwhile to focus on the trainee's knowledge and level of application of said model.

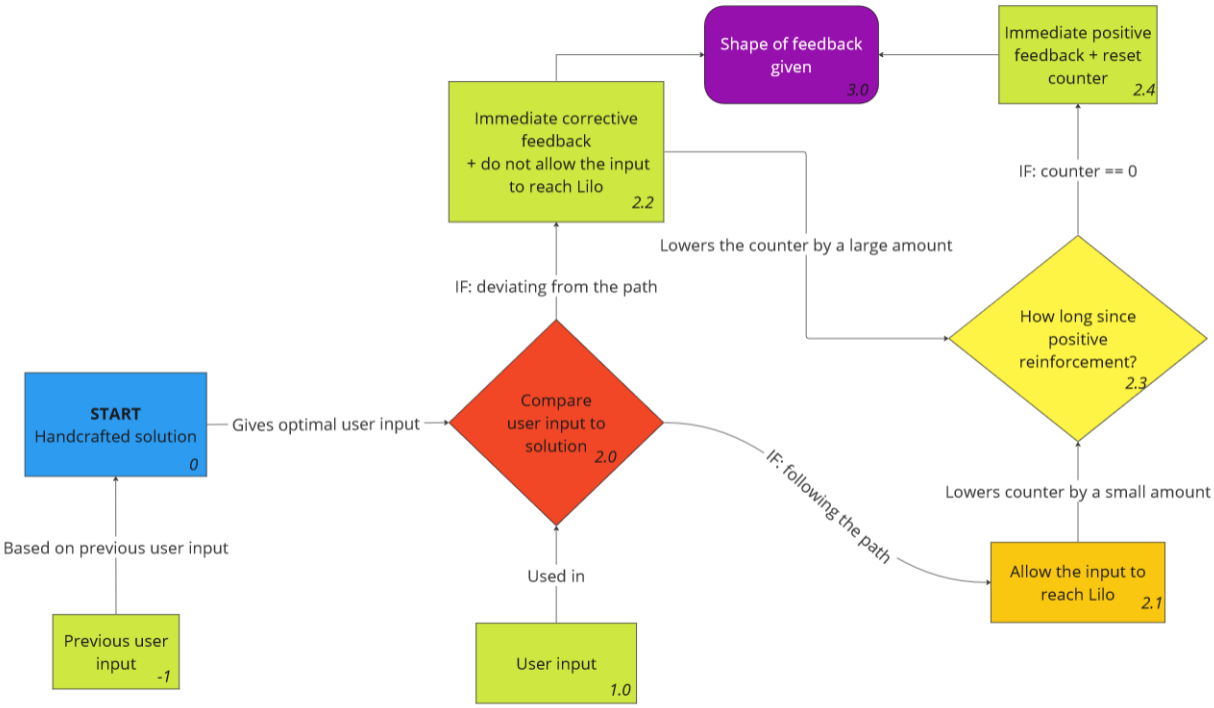
Design factors with possible design ideas for the future.



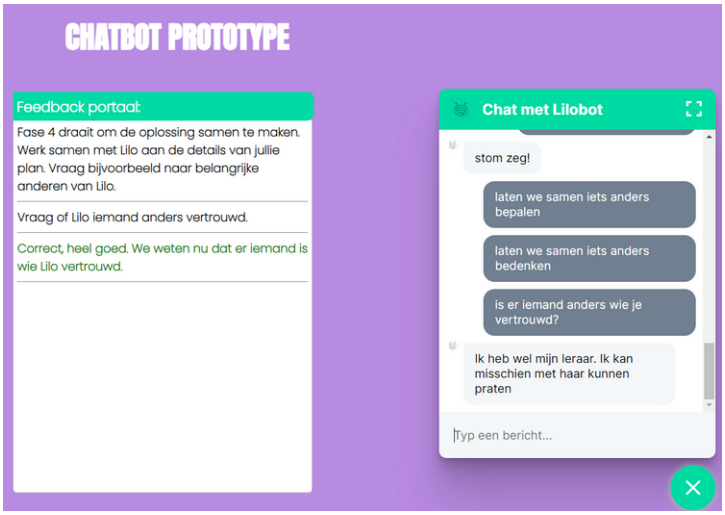
High-level overview of the system and user interactions.



Correct order of the desire, with possible side paths. Green denotes the correct path, while red shows an undesirable end.



High-level overview of how the feedback system will determine the correctness of the user input system.



Box 2.4 in execution. After a few erroneous inputs, the counter reaches 0 and the system gives positive feedback to the user.

How the feedback shows up for the user.

C. Experiment

Informed Consent Form
You are being invited to participate in an experiment titled: “Testing a feedback system for training purposes.” The experiment will take place between July and August. This experiment being done by Ayrton Armando Braam, and supervised by both Willem-Paul Brinkman and Mohammed Al Owayyed; All of which are affiliated with the TU Delft.

The purpose of this experiment is to gather the opinions of random users on a training system, which simulates a chat conversation with a virtual child. The system is meant to provide feedback for counselors in order to know how to perform the 5-phase model, which is a conversational technique developed by De Kindertelefoon. It will consist of two 20 minute sessions in which you engage with the simulation of a bullied child, after each session you will be asked to fill out a questionnaire. Though, the simulation of the child is not based upon a real story it is rooted in reality. The topics that can come up might be disturbing for some people, and we advise anyone who is sensitive to the topics of bullying, violence, and emotional distress to not participate in the experiment. You will be asked your give your age and gender, which will be categorized into groups for the data analysis. We will ask your opinion on the feedback you received during and after the session. We will ask you to fill out a questionnaire both prior and after the session, to compare and contrast in order to get the usefulness of the system. Both the questionnaires and the data we extrapolate from them will be shared for scientific purposes. Questions asked will cover your levels of self-efficacy and your thoughts on the usefulness of the system.

In general, data can be leaked. We will minimize any risks by getting your participation without registering identifying information. The private data will be stored privately and only accessible by the researchers. The anonymous questionnaires will be stored after the research has been concluded, it can be published in a public repository (e.g., 4TU.ResearchData).

Your participation in this experiment is entirely voluntary and you can withdraw at any time during the experiment session. The anonymous questionnaires, once posted, cannot be removed.

For more information please contact:
- Ayrton Armando Braam If you agree and consent to this Opening Statement, you can now fill in the consent form below.

Informed Consent Form		
Statements	**Yes**	**No**
I have read and understood the experiment information above. I have been able to ask questions about the experiment and my questions have been answered to my satisfaction.		
I consent voluntarily to be a participant in this experiment and understand that I can refuse to answer questions and withdraw from participation at any point during the experiment.		
I understand that I cannot withdraw from participation once the experiment ends.		
I understand that by participating in this experiment, that I will be faced with topics such as bullying, violence and emotional distress.		
I understand that taking part in the experiment involves giving my opinion on topics on the training system and how feedback within the system is given.		
I understand that taking part in the experiment involves the risk of possible data leakage. The researcher does everything to mitigate this risk by storing the collected data safely, and publish the anonymous data in a public repository.		
I understand that the information retrieved during participating in this experiment will be used for research and can be published in a scientific paper.		
I understand that my age and gender will be collected as part of the experiment and that it will it be part of the anonymous questionnaires.		
I agree that my anonymous opinions will be accessible for all purposes, including for example educational, research and commercial purposes.		
I give permission for the anonymous answers to the experiment that I provide to be archived in a public repository (e.g. 4TU Center for Research Data) so it can be used for future research and learning.		
By ticking this box, I agree to participate in this experiment.		

Figure 4: The informed consent form.

Questionnaire generalities

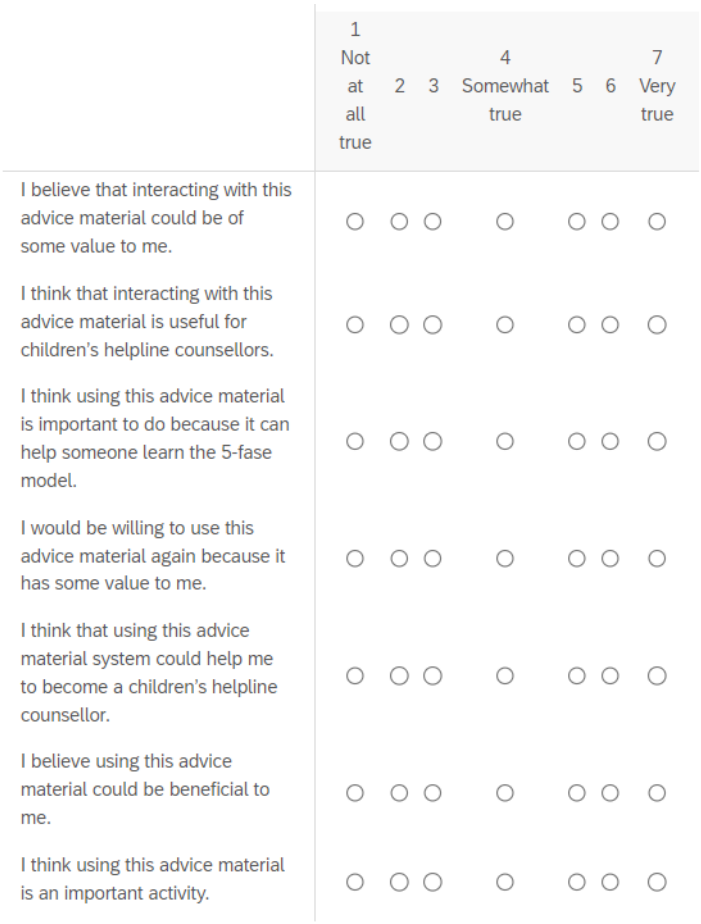
Here below is the general introduction to the questionnaires in Dutch as well as the IMI Likert scales for Perceived Usefulness and Enjoyment.

Welkom en bedankt voor het meedoen aan dit gedeelte van de questionnaires. Het doel van de questionnaire is data verzamelen voor een experiment waarin een prototype feedbacksysteem voor een virtual trainingsprogramma van een kinderhulplijn wordt gebruikt. Dit systeem geeft feedback aan de hand van een 5-fase model, bedoeld om mensen voor te bereiden op gesprekken met echte kinderen. De data verkregen met deze questionnaire wordt vergeleken met andere om te helpen met het beantwoorden van de research vragen van de master scriptie van Ayrton Armando Braam. Mocht je vragen hebben, neem alstublieft contact via whatsapp op +316 143 449 41.

Het onderwerp van de questionnaire is het 5-fase model. Het is een model bedoelt om counselors door een gesprek met een kind te leiden, terwijl de focus van het gesprek op het kind blijft. De vragen zijn gericht op het testen van de kennis die de mensen krijgen over het 5-fase model en om te zien wat

mensen nuttiger vinden tussen het feedbacksysteem en een alternatief.

De questionnaires zullen anoniem zijn, en geen persoonlijke informatie behalve geslacht en leeftijd zullen gevraagd worden. Geslacht en leeftijd is alleen nuttig on de data op te delen in verschillende demografen, die helpen met kijken of data van bepaalde groepen met elkaar overeenkomt, en wat dat zegt over het systeem. De questionnaires zullen samen niet meer dan 25 min duren.



	1 Not at all true	2	3	4 Somewhat true	5	6	7 Very true
I believe that interacting with this advice material could be of some value to me.	○	○	○	○	○	○	○
I think that interacting with this advice material is useful for children's helpline counsellors.	○	○	○	○	○	○	○
I think using this advice material is important to do because it can help someone learn the 5-fase model.	○	○	○	○	○	○	○
I would be willing to use this advice material again because it has some value to me.	○	○	○	○	○	○	○
I think that using this advice material system could help me to become a children's helpline counsellor.	○	○	○	○	○	○	○
I believe using this advice material could be beneficial to me.	○	○	○	○	○	○	○
I think using this advice material is an important activity.	○	○	○	○	○	○	○

The first part of the questionnaire, measuring the perceived usefulness of the advice material using the Intrinsic Motivation Inventory.

	1 Not at all true	2	3	4 Somewhat true	5	6	7 Very true
I enjoyed doing this activity very much.	○	○	○	○	○	○	○
This activity was fun to do.	○	○	○	○	○	○	○
I thought this activity was a boring activity.	○	○	○	○	○	○	○
This activity did not hold my attention at all.	○	○	○	○	○	○	○
I would describe this activity as very interesting.	○	○	○	○	○	○	○
I thought this activity was quite enjoyable.	○	○	○	○	○	○	○
While I was doing this activity, I was thinking about how much I enjoyed it.	○	○	○	○	○	○	○

The final part of the questionnaire, measuring the enjoyment participants had using the Intrinsic Motivation Inventory.

Closed questions
Below are tables showing the questions given to the participants of the experiment.

Questionnaire A				
Question	**A1**	**A2**	**A3**	**A4**
Wat is het doel van fase 1?	Achtergrond informatie van het kind verkrijgen.	De naam van het kind vragen.	Het kind informeren over hoe het gesprek verloopt.	Het kind een warm welkom geven.
Je hebt net het kind een warm welkom gegeven, wat gebeurt hierna?	Je vraagt naar de gewenste toestand van het kind.	Je legt aan het kind uit wat voor stappen ze kunnen nemen.	Je gaat luisteren naar het verhaal van het kind.	-
Wat kan een hulpverlener NIET doen in fase 2?	Vragen naar één specifieke situatie waarin het probleem zich voordoet.	De last van het kind erkennen.	De situatie concreet maken met behulp van doorvragen.	Vragen naar belangrijke andere van het kind en contact met ze opnemen.
Je hebt net aan het kind gevraagd wat hun wens is, wat is exact hiervoor gebeurd?	Je hebt het kind gekalmeerd.	Je hebt aan het kind uitgelegd wat voor stappen ze kunnen nemen	Je hebt geluistered naar het achtergrond verhaal van het kind	-
Wat is het doel van fase 3?	Samen met het kind een oplossing bedenken.	Het kind uitleggen wat zij kan doen.	De gewenste toestand van het kind is bepaald.	-
Wat is GEEN mogelijke stap van fase 4?	Vragen wat het kind zelf al geprobeerd heeft om het probleem op te lossen.	Vragen wat de gewenste toestand van het kind is.	Vragen naar belangrijke anderen van het kind.	-
Je hebt net de 2de fase, naar het kind luisteren, afgerond. Aan de start van de 3de fase, welke zin zou je tegen het kind kunnen zeggen?	Stel dat het pesten is verdwenen… hoe gaat jouw schooldag dan?	Wat heb je al geprobeerd waardoor je je een klein beetje vrolijk voelde op school?	Met wie kun je praten en wie luistert er naar jou?	Wanneer was de laatste keer dat je gepest werd?
Welke zin kan je in fase 5 tegen het kind zeggen?	Nou, dag he!	Hoe vond je het gesprek gaan? Welk cijfer zou je willen geven?	Ik ga het voor je regelen! Je hoort nog van me!	-
Jij en het kind hebben net afgesproken dat het kind met zijn ouders gaat praten, en hebt de 4de fase afgerond. Wat is het meest logische om nu te zeggen?	Ik stel voor om het gesprek nu te beëindigen zodat je aan de slag kunt gaan.	Beloof je dat je met je ouders zult praten?	Stel je een liniaal voor van 0 tot 10. 0 is je probleem op z'n allerergst, 10 is het opgelost. Waar sta je nu?	-

Figure 5: Closed Knowledge questions from questionnaire A

Questionnaire B				
Question	**A1**	**A2**	**A3**	**A4**
Welk van de volgende opties is een goede optie om contact met het kind te maken?	Het kind hun naam vragen.	Vragen naar de leeftijd van het kind.	Het kind geruststellen.	-
Wat zou een hulpverlener NIET moeten doen in fase 1?	De hulpverlener moet weten wat het doel en effect is van een warm welkom.	De hulpverlener vraagt naar de gewenste toestand van het kind.	De hulpverlener is in staat een warm welkom te geven en contact te maken aan het begin van het gesprek.	De hulpverlener vraagt het kind wat zijn behoefte is m.b.t. dit gesprek.
Wat is het doel van fase 2?	Luisteren naar het kind en informatie verzamelen.	Het kind geruststellen en vertellen dat het goed komt.	De wens van het kind bepalen en samen met het kind werken aan de wens.	Vragen naar belangrijke andere van het kind en contact met ze opnemen.
Wat zou een hulpverlener NIET moeten doen in fase 2?	De hulpverlener geeft juiste informatie of gaat samen met het kind op zoek naar passende informatie.	De hulpverlener stelt het kind gerust en vertelt dat alles goed komt.	De hulpverlener geeft erkenning voor de last die het kind ervaart van het probleem zodat het kind zich gehoord voelt.	De hulpverlener verheldert de last van een probleem aan de hand van een concrete situatie.
Wat is GEEN onderdeel van fase 3?	Samen met het kind een oplossing bedenken.	Onderzoek hoe de gewenste toestand voor het kind zou zijn.	De gewenste toestand van het kind bepalen.	-
Wat is het doel van fase 4?	Bouwen aan de gewenste toestand van het kind	Vragen naar belangrijke anderen van het kind.	Vragen naar sterke kanten het kind.	-
Welke van deze vragen is NIET logisch om aan het kind te stellen in de 4de fase?	Wat kun jij het allerbeste?	Wat vinden jouw vrienden/ouders/leraren leuk aan jou?	Met wie kun je praten en wie luistert er naar jou?	Wat hoop je dat er na dit gesprek anders zal zijn?
Je bent in fase 5 van het gesprek, met welke zin moet je het gesprek NIET afronden?	Heel veel succes! Je weet ons te vinden wanneer je daar behoefte aan hebt!	Dan ga ik nu de chat sluiten! Dank je wel voor het mooie gesprek!	Dan ga ik nu met je ouders bellen! Erg bedankt voor dit interessante gesprek!	-
Je bent nu bezig met het gesprek afronden, wat is exact hiervoor gebeurd?	Je hebt het kind verwelkomt	Je hebt met het kind gebouwd aan de gewenste toestand	Je hebt geluisterd naar het verhaal van het kind	-

Figure 6: Closed Knowledge Questions B

Open questions
Open questions given in the questionnaires.

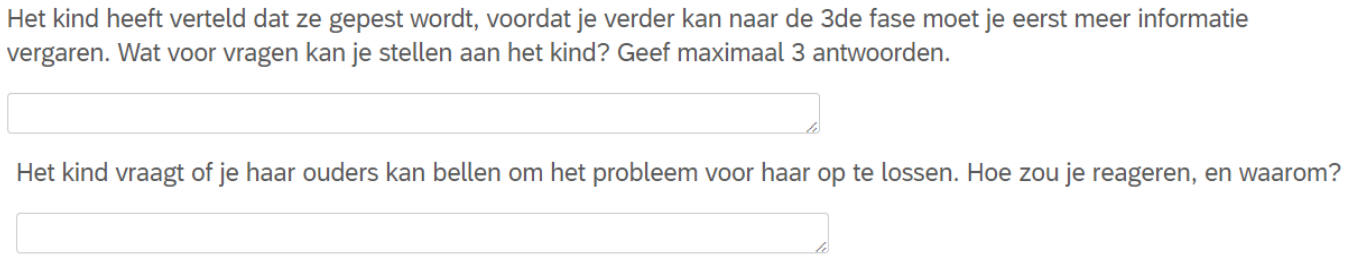


Figure 7: Open questions of the questionnaires.

D. additional scatter plots

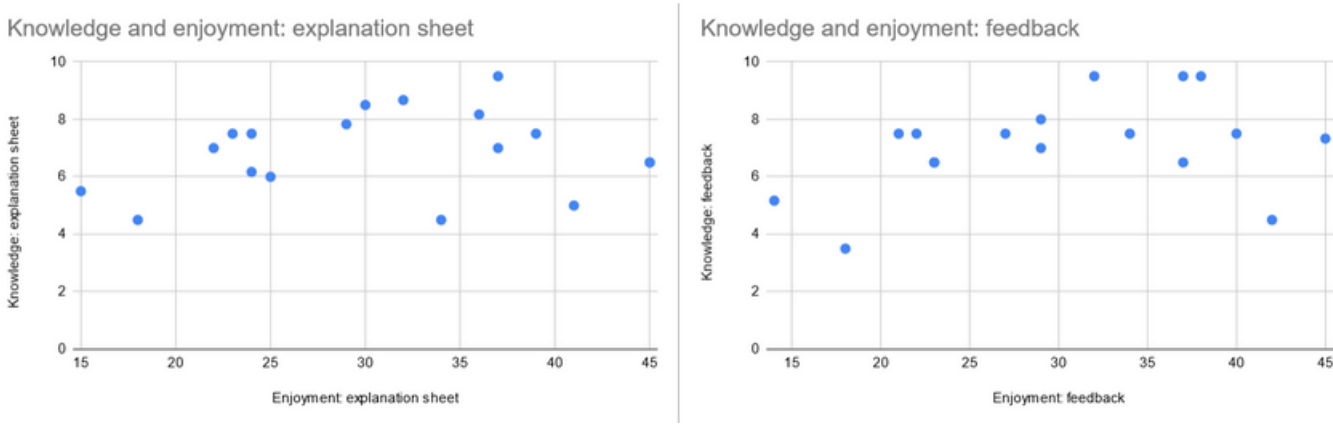


Figure 8: Two scatter plots of the knowledge and the enjoyment, for each condition

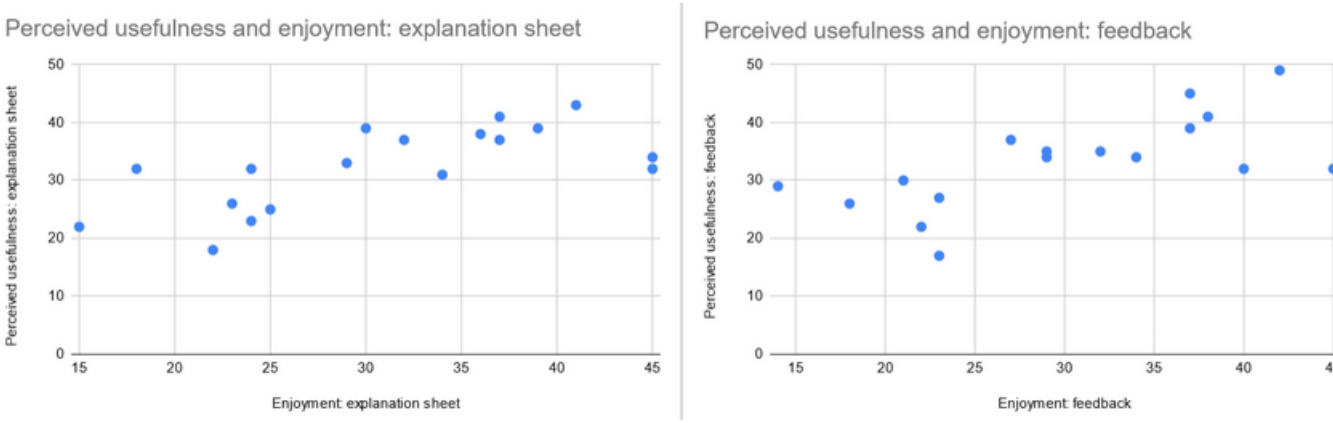


Figure 9: Two scatter plots of the perceived usefulness and the enjoyment, for each condition